С.А.Айвазян
И.С.Енюков
Л.Д.Мешалкин

ПРИКЛАДНАЯ СТАТИСТИКА

✦
✦

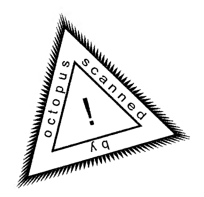
ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ

Справочное издание

Под редакцией
проф. С.А.Айвазяна



Москва
Финансы и статистика
1985



ББК 22.172
А11

Рецензенты *Е. Г. Ясин, А. И. Орлов*

Айвазян С. А. и др.
А11 Прикладная статистика: Исследование зависимостей: Справ. изд. / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин; Под ред. С. А. Айвазяна. — М.: Финансы и статистика, 1985. — 487 с., ил.
В пер. 1 р. 70 к. 13 000 экз.

Данная книга является логическим продолжением справочного издания «Прикладная статистика: Основы моделирования и первичная обработка данных», вышедшего в 1983 г. В ней рассматриваются методы корреляционного, регрессионного и дисперсионного анализа. Приводятся их алгоритмы и обзор программного обеспечения.

Для статистиков, экономистов, социологов, программистов.

А $\frac{1702060000-017}{010(01)-85}$ 66—84

ББК 22.172
517.8

ПРЕДИСЛОВИЕ

Вниманию читателя предлагается книга, продолжающая[1] реализацию замысла авторов: создать многотомное справочно-пособие по современным математическим методам статистической обработки данных, включающее в себя одновременное освещение необходимого *математического аппарата*, соответствующего *программного обеспечения* ЭВМ и рекомендаций по *преодолению вычислительных трудностей*, связанных с использованием описываемых методов и алгоритмов. Книга адресована специалистам различных сфер человеческой деятельности, использующим методы математической статистики и анализа данных в своей работе.

Для понимания материала книги читателю достаточно обладать математической подготовкой в объеме программ экономического или технического вуза либо ознакомиться с базовыми понятиями теории вероятностей и математической статистики, описанными в первом томе справочного издания [14]. В свою очередь освоение материала предлагаемой книги может служить надежной и удобной базой для более глубокого проникновения в предмет исследования, основанного на изучении специальных монографий и журнальных статей.

Тема книги, бесспорно, центральная во всем справочном издании. Она является таковой как по глубине и разнообразию разработанного к настоящему времени математического аппарата, так и по удельному весу использования описываемых методов и моделей в практических разработках разнообразного профиля.

---

[1] В 1983 г. вышла в свет книга: Айвазян С. А., Енюков И.С., Мешалкин Л. Д. Прикладная статистика: Основы моделирования и первичная обработка данных. — М.: Финансы и статистика. В ней, в частности, определена *прикладная статистика* как самостоятельная научная дисциплина, разрабатывающая и систематизирующая понятия, приемы, математические методы и модели, предназначенные для организации и обработки статистических данных с целью их удобного представления, интерпретации и получения научных и практических выводов (см. с. 19).

Главная цель, которую ставили перед собой авторы, — оснастить исследователя, использующего в своей работе статистические методы, инструментарием, необходимым для решения ключевой проблемы всякого исследования: как на основании частных результатов статистического наблюдения за анализируемыми событиями или показателями выявить и описать существующие между ними взаимосвязи. Именно эта проблема, *проблема статистического исследования зависимостей*, оказывается главной в решении таких типовых задач практики, как нормирование, прогноз, планирование, диагностика, оценка труднодоступных для непосредственного наблюдения и измерения характеристик анализируемой системы, оценка эффективности функционирования или качества объекта, регулирование параметров процесса или системы.

Авторы стремились к объективно сбалансированному представлению материала как по структуре книги, так и по ее содержанию. Однако широта и разноплановость затронутой проблемы не позволяют им претендовать на всеобъемлющий охват темы. Так, например, относительно узко представлена в данном томе тематика *статистического анализа динамических зависимостей*; не дано описания весьма полезного, в определенных типах задач, *аппарата логических решающих правил*[1]; не вошел в книгу материал, посвященный актуальной в прикладном плане (особенно в задачах управления технологическими процессами) тематике *планирования регрессионных экспериментов*.

Книга состоит из введения и четырех разделов.

Введение играет особую роль в понимании описываемых в дальнейшем методов и логики всей книги в целом. Можно сказать, что в нем в доступной для неискушенного читателя форме представлены содержание и логические связи всех частей книги. Приводятся основные постановки задач и «адреса» (в книге) их решения. Изложение проиллюстрировано простыми примерами. Поэтому сравнительно слабо подготовленному читателю рекомендуем не пожалеть времени на чтение введения.

Раздел I посвящен методам и приемам, позволяющим ответить на вопросы, *имеется ли вообще какая-либо связь между исследуемыми переменными, как измерить их тесноту и какова структура связей между показателями исследуемого набора?* При этом под структурой понимается характер всевозможных попарных двоичных взаимоотношений рассматривае-

[1] Читатель может познакомиться с этим аппаратом статистического исследования зависимостей, например, по книге [76].

мых признаков (по типу «связь есть» или «связи нет»), но не форма зависимости одного от другого. Методы, описанные в данном разделе, составляют содержание *корреляционного анализа*.

Раздел II содержит описание методов и моделей, позволяющих исследовать вид зависимости интересующего нас «выходного» (или «результирующего») количественного показателя от набора объясняющих переменных количественной природы (*регрессионный анализ*). В отдельной главе (гл. 12) рассмотрен случай, когда роль объясняющей переменной играет «время».

В разделе III решаются те же задачи, что и в разделе II, но в ситуации, когда в качестве объясняющих переменных выступают *неколичественные* или *одновременно неколичественные и количественные признаки* (*дисперсионный и ковариационный анализ*).

И наконец, в раздел IV включены глава, посвященная описанию методов статистического анализа так называемых *систем одновременных эконометрических уравнений* (т. е. набора одновременно выполняющихся соотношений, в которых одни и те же переменные могут участвовать в разных соотношениях: и в роли результирующего показателя, и в роли предсказывающей переменной), и глава, в которой дается обзор наиболее интересного отечественного и зарубежного *программного обеспечения* методов статистического исследования зависимостей.

Научная и педагогическая деятельность авторов, послужившая основой реализации предлагаемого издания, проводилась в Центральном экономико-математическом институте АН СССР, в Московском государственном университете им. М. В. Ломоносова и в Центральной научно-исследовательской лаборатории 4-го Главного управления при Министерстве здравоохранения СССР.

Книга написана: С. А. Айвазяном — предисловие, введение, гл. 1, 2, 5, 6, 11, выводы к гл. 9, введение и выводы к гл. 12, § 13.5 и приложение; Л. Д. Мешалкиным — гл. 3, 4, 7 (без § 7.5, 7.6 и п. 7.2.5), 10, 13 (без § 13.5); И. С. Енюковым — гл. 8, 15; В. В. Федоровым — гл. 9 (без § 9.6, 9.7 и п. 9.5.4), 12 (без введения и выводов), § 7.5 и 7.6; Ю. М. Кабановым — гл. 14 (без § 14.6); Е. З. Демиденко — п. 9.5.4; § 9.6, 9.7, 14.6; А. М. Шурыгиным — п. 7.2.5.

Авторы выражают глубокую благодарность А. И. Орлову и Е. Г. Ясину, взявшим на себя труд отрецензировать рукопись книги. Их критические замечания, бесспорно, способствовали повышению качества данного издания. Авторы при-

знательны также В. Н. Вапнику, предоставившему им материалы для написания п. 6.3.1, а также А. Б. Успенскому, Е. З. Демиденко, А. М. Шурыгину, Арк. И. Верескову и О. В. Лепскому, участвовавшим в обсуждении отдельных частей рукописи, а также и Л. Ю. Метт, вложившей большой труд в оформление рукописи.

Положительную роль в замысле и содержании книги сыграли постоянные контакты авторов со своими коллегами по научному семинару «Многомерный статистический анализ и вероятностное моделирование реальных процессов» (действующему в рамках Научного совета АН СССР по комплексной проблеме «Оптимальное планирование и управление народным хозяйством» и Совета по автоматизации научных исследований при Президиуме АН СССР), а также по Всесоюзному научно-методическому семинару «Вычислительные вопросы математической статистики», действующему в Московском государственном университете им. М. В. Ломоносова под руководством Ю. В. Прохорова.

С. А. Айвазян

Введение. СТАТИСТИЧЕСКОЕ ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ СОДЕРЖАНИЕ, ЗАДАЧИ, ОБЛАСТИ ПРИМЕНЕНИЯ

В.1. Предварительное обсуждение задач

Любой закон природы или общественного развития может быть выражен в конечном счете в виде описания характера или структуры взаимосвязей (зависимостей), существующих между изучаемыми явлениями или показателями (переменными величинами или просто *переменными*). Если эти зависимости: а) *стохастичны* по своей природе, т. е. позволяют устанавливать лишь вероятностные логические соотношения между изучаемыми событиями A и B, а именно соотношения типа «из факта осуществления события A следует, что событие B должно произойти, но не обязательно, а лишь с некоторой (как правило, близкой к единице) вероятностью P»; б) выявляются на основании *статистического наблюдения* за анализируемыми событиями или переменными, осуществляемого по выборке из интересующей нас генеральной совокупности [14, п. 5.4.2], то мы оказываемся в рамках проблемы *статистического исследования зависимостей*. Соответствующий математический аппарат, будучи таким образом нацеленным в первую очередь на решение основной проблемы естествознания: как по отдельным, частным наблюдениям выявить и описать интересующую нас общую закономерность? — занимает, бесспорно, центральное место во всем прикладном математическом анализе.

Перед тем как перейти к формулировке общей и частных задач статистического исследования зависимостей, условимся описывать функционирование изучаемого реального объекта (системы, процесса, явления) набором переменных (рис. В.1), среди которых:

$x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ — так называемые *«входные» переменные*, описывающие *условия функционирования* (часть из них, как правило, поддается регулированию или частичному управлению); в соответствующих математических моделях их называют независимыми, факторами-аргументами, экзогенными, предикторными (или просто предикторами, т. е. предсказателями), объясняющими (в книге мы будем использовать в основном два последних термина);

$y^{(1)}, y^{(2)}, \ldots, y^{(m)}$ — *выходные* переменные, характеризующие *поведение или результат* (*эффективность*) *функционирования*; в математических моделях их называют зависимыми, откликами, эндогенными, результирующими или объясняемыми (в книге используются в основном два последних термина);

$\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(m)}$ — латентные (т. е. скрытые, не поддающиеся непосредственному измерению) *случайные «остаточные» компоненты*, отражающие влияние (соответственно на

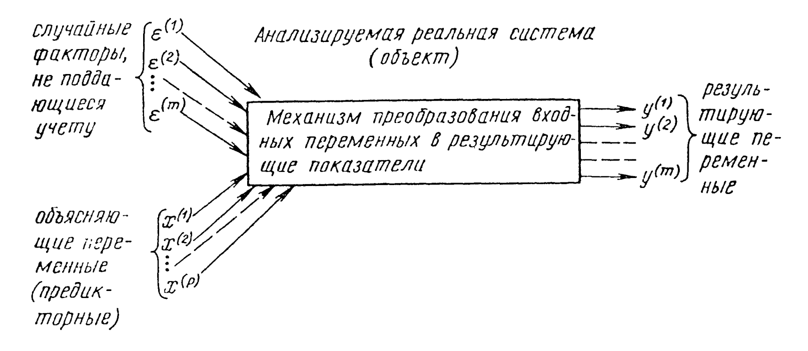


Рис. В.1. Общая схема взаимодействия переменных при статистическом исследовании зависимостей

$y^{(1)}, y^{(2)}, \ldots, y^{(m)}$) неучтенных «на входе» факторов, а также случайные ошибки в измерении анализируемых показателей (в математических моделях мы их, как правило, будем именовать просто «*остатками*»).

Тогда общая задача статистического исследования зависимостей (в терминах изучаемых показателей) может быть сформулирована следующим образом:

по результатам n измерений

$$\left\{\left(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; \; y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(m)}\right)\right\}_{i=1,2,\ldots,n} \qquad (\text{B.1})$$

исследуемых переменных на объектах (*системах, процессах*) *анализируемой совокупности построить такую* (*векторнозначную*) *функцию*

$$\mathbf{f}(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \begin{pmatrix} f^{(1)}(x^{(1)}, \ldots, x^{(p)}) \\ f^{(2)}(x^{(1)}, \ldots, x^{(p)}) \\ \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \; \cdot \\ f^{(m)}(x^{(1)}, \ldots, x^{(p)}) \end{pmatrix} \qquad (\text{B.2})$$

которая позволила бы наилучшим (в определенном смысле) образом восстанавливать значения результирующих (прогнозируемых) переменных $Y = (y^{(1)}, y^{(2)}, \ldots, y^{(m)})'$ *по заданным значениям объясняющих (предикторных) переменных* $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})'$[1].

Данная формулировка задачи нуждается в уточнениях. В частности, прежде всего мы должны ответить на следующие вопросы:

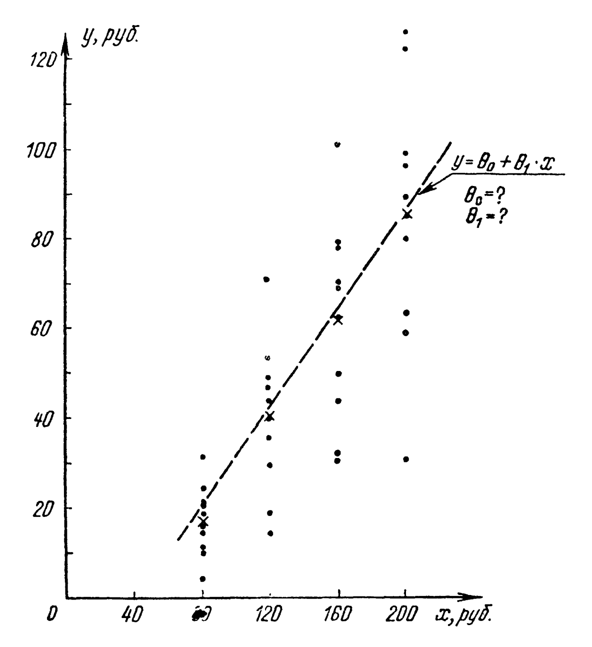


Рис. В.2. Графическое представление результатов обследования 40 семей по их среднедушевому доходу (x_i) и среднедушевым денежным сбережениям (y_i)

а) каково математическое выражение (или *структура модели* [14, с. 68—73]) искомой зависимости между Y и X, записанное в терминах Y, X, $\mathbf{f}(X)$ и $\varepsilon = (\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(m)})$?

б) в соответствии с каким именно *критерием качества аппроксимации* значений Y с помощью функции $\mathbf{f}(X)$ мы будем

---

[1]Здесь и далее штрих при векторе или матрице означает операцию их транспонирования. В данном случае это означает, что Y и X — соответственно m- и p-мерные вектор-столбцы.

определять наилучший способ восстановления значений результирующих показателей по заданным значениям объясняющих переменных?

в) с какой именно *прикладной целью* мы проводим все наше исследование, т. е. для решения каких конкретных задач мы собираемся использовать построенную в результате исследования функцию $\mathfrak{f}(X)$?

Прежде чем обсуждать эти вопросы, рассмотрим пример.

Пример В.1. Анализируется «поведение» двумерной случайной величины (ξ, η), где ξ (руб.) — среднедушевой доход и η (руб.) — среднедушевые денежные сбережения в семье, случайно извлеченной из рассматриваемой совокупности семей, однородной по своему потребительскому поведению (см., например, [128]). В табл. В.1 и на рис. В.2 представлены исходные статистические данные вида (В.1), характеризующие среднедушевые величины дохода (x_i, руб.) и денежных сбережений (y_i, руб.) за определенный отрезок времени, а именно за месяц, в каждой (i-й, $i = 1, 2, \ldots, n$) обследованной семье рассматриваемой совокупности семей (в данном условном примере объем n статистически обследованной совокупности семей равнялся 40). В этом примере имелась возможность при отборе исходных данных (выборки) *контролировать значения предикторной переменной* ξ (условия активного эксперимента [14, с. 121]), что позволило, в частности, разбить статистически обследованные семьи на четыре равные по объему группы по доходам.

Мы видим, что даже в пределах каждой из этих групп величины среднедушевых сбережений семей подвержены некоторому неконтролируемому разбросу, обусловленному влиянием множества не поддающихся строгому учету и контролю факторов (т. е. налицо упомянутый выше *стохастический характер* зависимости между x и y). Однако это еще не значит, что расположение точек (x_i, y_i), являющихся геометрическим изображением результатов обследования семей по доходу и сбережениям, должно быть совершенно хаотичным и не должно обнаруживать некоторой *вполне определенной тенденции*, характеризующей зависимость денежных сбережений в семье (η) от ее среднедушевого дохода (ξ). При исследовании подобных зависимостей встают следующие основные вопросы (в скобках после вопроса указываются главы, параграфы или пункты настоящей книги, ему посвященные).

1. Как исходя из конкретных прикладных целей исследования определить смысл, в котором понимается исследуемая зависимость? (В.2, § 5.3.)

2. Имеется ли вообще какая-либо связь между исследуе-

мыми переменными (а в случае многих переменных — какова структура этих связей?) и как измерить тесноту этой связи? (Гл. 1—4.)

3. Каков *общий* математический вид искомой связи между η и ξ, т. е. как определяется *общая структура* соответствующей математической модели? (Гл. 6.)

4. Как, отправляясь от принятой общей структуры модели, провести необходимую вычислительную обработку исходных данных (B.1) с целью получения *конкретного вида* зависимости η от ξ, что позволит в данном случае производить количественную оценку неизвестных денежных сбережений семьи по заданной величине ее среднедушевого дохода? (Гл. 7—10, 13, 14.)

5. Поскольку наши выводы основаны на обработке *ограниченного ряда* наблюдений, то их количественные характеристики, естественно, подвержены (при повторениях соответствующих выборочных обследований) некоторому случайному разбросу. Как оценить *степень точности наших выводов*? (Гл. 11.)

6. Как решать все вопросы в ситуациях, когда среди объясняющих (предикторных) переменных могут быть и неколичественные? (Гл. 13.)

7. И наконец, если при сборе исходной статистической информации мы находимся в условиях *активного эксперимента* [14, с. 12], то как, при заданных затратах на наблюдения, оптимально выбрать матрицу плана [14, с. 26, 68], т. е. как определить те значения объясняющих (предикторных) переменных и то распределение заданного общего числа наблюдений между этими значениями, которые являются в некотором смысле наиболее выгодными с точки зрения достижения наивысшей точности наших статистических выводов?

Вернемся к нашему примеру и попробуем ответить на некоторые из поставленных здесь вопросов, в том числе на принципиальные вопросы а), б) и в), ответы на которые позволяют уточнить общую формулировку задачи статистического исследования зависимостей, данную выше.

Начнем «с конца», т. е. с уточнения *конечных прикладных целей исследования* (см. вопросы 1, а также а) и в)). Известно, что из двух анализируемых характеристик материальной состоятельности семьи характеристика денежных сбережений (η) относится к категории статистически труднодоступных: содержащиеся в ежегодных и единовременных выборочных семейных бюджетных обследованиях ЦСУ [85] сведения о сбережениях, как правило, непредставительны. Поэтому главной конечной целью нашего исследования (опирающегося, как мы

Среднедушевой доход, руб.	$x_1 = x_2 = \ldots = x_{10} = x_1^0 = 80$	$x_{11} = x_{12} = \ldots = x_{20} = x_2^0 = 120$
Среднедушевые сбережения η	$y_1 = 15{,}2$ $y_2 = 10{,}7$ $y_3 = 18{,}5$ $y_4 = 14{,}9$ $y_5 = 24{,}1$ $y_6 = 10{,}3$ $y_7 = 14{,}2$ $y_8 = 31{,}0$ $y_9 = 20{,}4$ $y_{10} = 20{,}0$	$y_{11} = 70{,}1$ $y_{12} = 35{,}0$ $y_{13} = 43{,}0$ $y_{14} = 29{,}0$ $y_{15} = 17{,}0$ $y_{16} = 48{,}2$ $y_{17} = 18{,}9$ $y_{18} = 53{,}0$ $y_{19} = 39{,}4$ $y_{20} = 46{,}2$
Средние сбережения для семей данной группы	$\bar{y}(x_1^0) = \frac{1}{10}\sum_{i=1}^{10} y_i = 17{,}9$	$\bar{y}(x_2^0) = \frac{1}{10}\sum_{i=11}^{20} y_i = 40{,}0$
Среднеквадратическое отклонение s и коэффициент вариации $\hat{V}$ сбережений для семей данной группы по доходам	$s(x_1^0) = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(y_i - \bar{y}(x_1^0))^2} = 6{,}4$. $\hat{V}(x_1^0) = 36\ \%$	$s(x_2^0) = \sqrt{\frac{1}{9}\sum_{i=11}^{20}(y_i - \bar{y}(x_2^0))^2} = 16{,}0$ $\hat{V}(x_2^0) = 40\ \%$

будем всегда предполагать, на достоверную и репрезентативную выборку исходных данных) является возможность восстановления (прогноза):

удельной (т. е. в расчете на одного члена семьи за определенный отрезок времени) величины денежных сбережений в *конкретной семье* ($y(x)$) по заданному значению ее среднедушевого дохода x;

удельной величины *средних* денежных сбережений ($y_{ср}(x)$) в семьях данной группы x по доходам.

Таблица В.1

Среднедушевой доход, руб.	$x_{21}=x_{22}=\ldots=x_{30}=x_3^0=160$	$x_{31}=x_{32}=\ldots=x_{40}=x_4^0=200$
Среднедушевые сбережения η	$y_{21}=49,6$ $y_{22}=69,4$ $y_{23}=77,8$ $y_{24}=43,0$ $y_{25}=31,8$ $y_{26}=62,6$ $y_{27}=100,2$ $y_{28}=68,8$ $y_{29}=78,0$ $y_{30}=29,6$	$y_{31}=125,5$ $y_{32}=88,3$ $y_{33}=62,0$ $y_{34}=58,8$ $y_{35}=84,0$ $y_{36}=79,0$ $y_{37}=95,5$ $y_{38}=120,8$ $y_{39}=98,1$ $y_{40}=29,7$
Средние сбережения для семей данной группы	$\bar{y}(x_3^0)=\frac{1}{10}\sum_{i=21}^{30} y_i=61,1$	$\bar{y}(x_4^0)=\frac{1}{10}\sum_{i=31}^{40} y_i=84,2$
Среднеквадратическое отклонение s и коэффициент вариации $\widehat{V}$ сбережений для семей данной группы по доходам	$s(x_3^0)=\sqrt{\frac{1}{9}\sum_{i=21}^{30}(y_i-\bar{y}(x_3^0))^2}=22,6$ $\widehat{V}(x_3^0)=37\ \%$	$s(x_4^0)=\sqrt{\frac{1}{9}\sum_{i=31}^{40}(y_i-\bar{y}(x_4^0))^2}=28,9$ $\widehat{V}(x_4^0)=34\ \%$

Этой цели мы сможем достигнуть, если сумеем математически описать закономерность изменения условных теоретических средних значений $y_{\text{ср}}(x)=\mathbf{E}(\eta\mid\xi=x)$[1] в зависимости

[1]Здесь и далее используются терминология и обозначения [14]. В частности, знаком **E** обозначается операция теоретического осреднения, а знаком **D** — операция вычисления дисперсии случайных величин, стоящих за ними. Вертикальная черта разделяет случайную величину, над которой производится операция осреднения или вычисления дисперсии, и условие, при котором эта операция производится.

от x, а также изучить характер случайного разброса денежных сбережений $y(x)$ *отдельных* семей данной группы x по доходам относительно своего среднего значения $y_{ср}(x)$ (при любом интересующем нас значении среднедушевого дохода x). Это естественным образом приводит нас к необходимости рассмотрения математической модели вида

$$\eta = f(x) + \varepsilon, \qquad (В.3)$$

в которой остаточная компонента ε отражает случайное отклонение денежных сбережений наугад выбранной отдельной семьи с доходом $\xi = x$ от среднего значения $y_{ср}(x) = \mathbf{E}(\eta \mid \xi = x)$ этих сбережений, подсчитанного по всем семьям данной группы по доходам, а функция $f(x)$ описывает характер изменения условного среднего $y_{ср}(x)$ (при $\xi = x$) в зависимости от изменения x, если дополнительно прийти к соглашению, что характер случайного разброса величин $y(x) = (\eta \mid \xi = x)$ относительно своих средних $y_{ср}(x)$ таков, что $\mathbf{E}(\varepsilon \mid \xi = x) = 0$ при всех x.

Таким образом, из (В.3) мы непосредственно получаем

$$y_{ср}(x) = \mathbf{E}(\eta \mid \xi = x) = f(x). \qquad (В.4)$$

Чтобы покончить с вопросами 1, а) и в), остается уточнить общую структуру модели, т. е. определить, в *каком классе* $\mathbf{F}$ *функций* $f(x)$ *мы будем производить аппроксимацию искомой зависимости* $y_{ср}(x)$.

В нашем случае, учитывая однородный (по характеру потребительского поведения) состав исследуемой совокупности семей, естественно исходить из гипотезы об одинаковой (в среднем) склонности семей к сбережениям, выражающейся, в частности, в том, что все семьи начиная с некоторого «порогового» уровня дохода, склонны отделять в сбережения в среднем одинаковую долю дохода. Математически, как легко понять, это выразится в виде

$$y_{ср}(x) = \theta_0 + \theta_1 x, \qquad (В.5)$$

где θ_0 и θ_1 — некоторые константы (*неизвестные параметры модели*). Так что

$$\mathbf{F} = \{\theta_0 + \theta_1 x\}, \qquad (В.6)$$

где под $\{f(x; \Theta)\}$ понимается *семейство* всех тех функций $f(x; \Theta)$, которые могут быть получены при подстановке вместо Θ ее различных конкретных значений (Θ — векторный параметр).

Такой выбор «класса допустимых решений» $\mathbf{F} = \{f(x)\}$ подтверждается и характером расположения совокупности то-

чек, являющихся геометрическим изображением исходных данных в нашем примере (см. на рис. В.2 расположение «крестиков», ординаты которых определяются экспериментально подсчитанными, т. е. вычисленными на основании имеющихся выборочных данных, условными средними $\bar{y}(x_i^0)$, $i = 1, 2, 3, 4$)[1].

И наконец, следует уточнить, *в соответствии с каким именно критерием качества аппроксимации* неизвестных величин среднедушевых семейных денежных сбережений $y(x)$ и $y_{ср}(x)$ с помощью функции $\theta_0 + \theta_1 x$ мы будем определять наилучший способ прогноза $y_{ср}(x)$ по x. Наиболее обоснованное и точное решение этого вопроса опирается на знание вероятностной природы (а именно типа закона распределения вероятностей) остатков ε в модели (В.3). Так, например, известно [14, с. 281], что если предположить, что при любых значениях x распределение вероятностей остатков ε описывается $(0, \sigma^2)$-нормальным законом (т. е. нормальным законом со средним значением, равным нулю, и с некоторой, вообще говоря, неизвестной, но *постоянной*, т. е. не зависящей от x дисперсией σ^2) и что остатки $\varepsilon(x_i)$, $i = 1, 2, \ldots, n$, характеризующие различные наблюдения, статистически независимы, то наименьшая ошибка прогноза $y_{ср}(x)$ с помощью модели $f(x) \in \mathbf{F}$ (т. е. функция $f(x)$ подбирается из класса $\mathbf{F}$) обеспечивается требованием метода наименьших квадратов

$$\Delta_n(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \to \min_{f \in \mathbf{F}}. \quad \text{(B.7)}$$

В нашем примере явно нарушено условие постоянства дисперсии остатков (см. табл. В.1), т. е. условная дисперсия $\mathbf{D}(\varepsilon \mid \xi = x) = \mathbf{D}(\eta - \theta_0 - \theta_1 \cdot \xi \mid \xi = x) = \sigma^2(x)$ существенно зависит от значения x. Можно устранить это нарушение, поделив все анализируемые величины, откладываемые по оси η, а следовательно, и остатки $\varepsilon(x)$, на значения $s(x)$ (являющиеся статистическими оценками для $\sigma(x)$), т. е. перейдя к анализу остатков $\tilde{\varepsilon}(x) = \varepsilon(x)/s(x)$. Тогда можно показать (с помощью методов, описанных,

[1]Обращаем внимание читателя на разницу в смысле и обозначениях экспериментальных (выборочных) и теоретических условных средних соответственно $\bar{y}(x)$ и $y_{ср}(x)$. Строго говоря, на практике теоретических средних мы никогда знать не можем, однако мы опираемся в своем исследовании на тот факт, что в соответствии с законом больших чисел [14, с. 231] $\bar{y}(x) \to y_{ср}(x)$ (по вероятности), когда число наблюдений, по которым подсчитано $\bar{y}(x)$, стремится к бесконечности.

например, в [14, § 11.1]), что гипотеза о (0; σ^2)-нормальном характере распределения остатков $\tilde{\varepsilon}(x)$ не противоречит имеющимся в нашем распоряжении данным (представленным в табл. В.1) и, следовательно, требование (В.7) приводит к необходимости решения экстремальной задачи вида

$$\Delta_n(f) = \Delta_n(\theta_0, \theta_1) = \sum_{i=1}^{n} \left(\frac{y_i - \theta_0 - \theta_1 x_i}{s(x_i)} \right)^2 \to \min_{\theta_0, \theta_1}, \qquad (В.7')$$

т. е. к системе из двух линейных уравнений с двумя неизвестными (θ_0 и θ_1):

$$\begin{aligned} \frac{\partial \Delta_n(\theta_0, \theta_1)}{\partial \theta_0} &= -2 \sum_{i=1}^{n} s^{-2}(x_i) \cdot (y_i - \theta_0 - \theta_1 x_i) = 0; \\ \frac{\partial \Delta_n(\theta_0, \theta_1)}{\partial \theta_1} &= -2 \sum_{i=1}^{n} s^{-2}(x_i) \cdot x_i \cdot (y_i - \theta_0 - \theta_1 x_i) = 0. \end{aligned} \qquad (В.7'')$$

Решение системы (В.7″) дает нам в качестве оценок $\widehat{\theta_0}$ и $\widehat{\theta_1}$ для неизвестных параметров соответственно θ_0 и θ_1 выражения:

$$\widehat{\theta_1} = \frac{\left(\sum_{i=1}^{n} s^{-2}(x_i)\right)\left(\sum_{i=1}^{n} s^{-2}(x_i) \cdot x_i y_i\right) - \left(\sum_{i=1}^{n} s^{-2}(x_i) \cdot x_i\right) \times \left(\sum_{i=1}^{n} s^{-2}(x_i) \cdot y_i\right)}{\left(\sum_{i=1}^{n} s^{-2}(x_i)\right)\left(\sum_{i=1}^{n} s^{-2}(x_i) \cdot x_i^2\right) - \left(\sum_{i=1}^{n} s^{-2}(x_i) \cdot x_i\right)^2};$$

$$\widehat{\theta_0} = \frac{\sum_{i=1}^{n} s^{-2}(x_i) \cdot y_i}{\sum_{i=1}^{n} s^{-2}(x_i)} - \widehat{\theta_1} \cdot \frac{\sum_{i=1}^{n} s^{-2}(x_i) \cdot x_i}{\sum_{i=1}^{n} s^{-2}(x_i)}.$$

Расчет по этим формулам с использованием данных табл. В.1 дает нам решение задачи 4:

$\widehat{\theta_1} = 0{,}685;$

$\widehat{\theta_0} = -40{,}360,$

так что статистическая оценка искомой зависимости средней величины среднедушевых семейных сбережений $y_{\text{ср}}(x)$ от значения среднедушевого дохода семей данной доходной группы x имеет в этом случае вид

$$\widehat{y_{\text{ср}}}(x) = -40{,}36 + 0{,}685 \cdot x.$$

При другой статистической природе остатков ε или при отсутствии достаточной информации о типе их вероятностного распределения возможен иной, чем по (В.7), выбор критерия качества аппроксимации Δ_n (см. гл. 7). Отметим, однако, что наиболее широкое распространение в статистической практике именно критерия наименьших квадратов (В.7) подкреплено рядом исследований [15, 196]. В них обосновываются хорошие прогностические свойства моделей, полученных в соответствии с (В.7) и в ситуациях, характеризующихся различными отклонениями от нормальности и взаимной независимости остатков $\varepsilon(x)$.

Заканчивая обсуждение примера В.1 и возвращаясь к общему описанию задач статистического исследования зависимостей, отметим, что функции $\mathbf{f}(X) = \mathbf{E}(\eta \mid \xi = X)$, описывающие поведение условных средних результирующего показателя η (вычисленных при значениях предикторных переменных ξ, зафиксированных на уровне $\xi = X$) в зависимости от изменения X, принято называть *функциями регрессии* (подробнее о различных определениях функции регрессии см. в гл. 5).

В.2. Какова конечная прикладная цель статистического исследования зависимостей?

С этого вопроса должно начинаться любое статистическое исследование зависимостей[1]. Ведь от ответа на этот вопрос существенно зависят план исследования, выбор общей структуры математической модели, интерпретация получаемых статистических характеристик и выводов и т. д.

---

[1]Опыт вынуждает констатировать наличие большого числа прикладных исследовательских работ (статей, диссертаций, научных отчетов и т. д.), в которых этот тезис, казалось бы, тривиальный и очевидно справедливый, предается забвению. В подобных работах строятся различные модели, проводится большое число вычислений, анализируются статистические свойства полученных характеристик и т. п., но все это в конечном счете как бы «повисает в воздухе», вызывает у компетентного читателя вопросы: «ну и что?» или «зачем это нужно?», поскольку остается неясным, *как и для решения каких именно конкретных прикладных задач* предполагается использовать результаты проделанных математических упражнений

Итак, для чего же строятся математические модели типа (В.3), описывающие статистические зависимости между исследуемыми переменными: результирующими показателями $Y = (y^{(1)}, y^{(2)}, \ldots, y^{(m)})$, с одной стороны, и соответствующими объясняющими (предикторными) переменными $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$, с другой стороны?

Выделим три основных типа конечных прикладных целей подобных исследований, расположив их как бы по нарастанию глубины проникновения в содержательную сущность анализируемой конкретной задачи.

Тип 1: *Установление самого факта наличия (или отсутствия) статистически значимой связи между Y и X.* При такой постановке задачи статистический вывод имеет двоичную (альтернативную) природу — «связь есть» или «связи нет» — и сопровождается обычно лишь численной характеристикой (измерителем) степени тесноты исследуемой зависимости. Выбор *формы связи* (т. е. класса допустимых решений $\mathbf{F}$ и конкретного вида функции $\mathbf{f}(X)$ в модели (В.3)) и состава предикторов X играет подчиненную роль и нацелен исключительно на максимизацию величины этого измерителя степени тесноты связи: исследователю часто не приходится даже «добираться» до конкретного вида функции $\mathbf{f}(X)$ и тем более он не претендует на анализ причинных влияний переменных X на результирующие показатели.

Тип 2: *прогноз (восстановление) неизвестных значений интересующих нас индивидуальных ($Y(X) = (\eta \mid \xi = X)$) или средних ($Y_{ср}(X) = \mathbf{E}(\eta \mid \xi = X)$) значений исследуемых результирующих показателей по заданным значениям X соответствующих (предикторных) переменных.* При такой постановке задачи статистический вывод включает в себя описание интервала (области) $A_p(X)$ вероятных значений прогнозируемого показателя $Y_{ср}(X)$ или $Y(X)$ и сопровождается величиной доверительной вероятности P, с которой гарантируется справедливость нашего прогноза, формализуемого с помощью утверждения вида $\{Y(X) \in A_p(X)\}$ или $\{Y_{ср}(X) \in A_p(X)\}$. Как и в предыдущем случае, выбор формы связи (т. е. класса допустимых решений $\mathbf{F}$ и конкретного вида функции $\mathbf{f}(X)$ в модели (В.3)) и состава предикторов X играет подчиненную роль и нацелен исключительно на минимизацию ошибки получаемого прогноза. Однако в данном случае (в отличие от предыдущего) исследователь *существенно использует* значения функции $\mathbf{f}(X)$, которые являются отправной точкой при построении прогнозных интервалов (областей) $A_p(X)$. Последние обычно определяются в форме множества всех тех значений Y, которые удовлетворяют неравенствам

$$\mathbf{f}(X) - \varepsilon_p(X, n) \leqslant Y \leqslant \mathbf{f}(X) + \varepsilon_p(X, n), \quad \text{(B.8)}$$

где $\varepsilon_p(X, n)$ — гарантируемая (с вероятностью не меньшей заданного значения P) максимальная величина ошибки прогноза[1]. Таким образом, исследователя интересуют в данном случае *лишь значения* функции $\mathbf{f}(X)$, но *не ее структура*, определяющая, в частности, соотношение удельных весов влияния объясняющих переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ на каждый из результирующих показателей $y^{(k)}$ $(k = 1, 2, \ldots, m)$. Так, например, если при статистическом оценивании неизвестной истинной зависимости

$$f(X) = y_{\text{ср}}(x^{(1)}, x^{(2)}) = 1 + 3x^{(1)} + 5x^{(2)} \quad \text{(B.9)}$$

исследователю удалось получить оценку функции $\mathbf{f}(X)$ в виде

$$\hat{f}(X) = 1 + 6x^{(1)} - x^{(2)} \quad \text{(B.9')}$$

и при этом было установлено, что объясняющие переменные $x^{(1)}$ и $x^{(2)}$ связаны между собой *«почти функциональной»* линейной зависимостью[2]

$$x^{(1)} \approx 2x^{(2)}, \quad \text{(B.10)}$$

то функция $\hat{f}(X)$ будет обладать хорошими прогностическими свойствами, несмотря на существенное отличие ее коэффициентов при $x^{(1)}$ и $x^{(2)}$ от соответствующих коэффициентов истинной функции $f(X)$. (Обращаем внимание читателя на тот факт, что коэффициенты при $x^{(2)}$ в функциях $f(X)$ и $\hat{f}(X)$ отличаются даже по знаку!) При подстановке заданных значений объясняющих переменных $x^{(1)}$ и $x^{(2)}$ в правые части (B.9) и (B.9′), при условии, что эти значения связаны приближенным соотношением (B.10), мы будем получать совпадающие (или приближенно совпадающие) результаты $f(X)$ и $\hat{f}(X)$, характеризующие усредненную величину $y_{\text{ср}}(X)$ исследуемого результирующего показателя.

Тип 3: *выявление причинных связей между объясняющими переменными X и результирующими показателями Y, частич-*

---

[1]Напоминаем читателю, что $\mathbf{f}$, ε и Y являются m-мерными векторами (см. (B.2)), так что запись (B.8) означает справедливость m соответствующих покомпонентных неравенств.

[2]Говоря о «почти функциональной» линейной зависимости между $x^{(1)}$ и $x^{(2)}$, мы имеем в виду близость к единице (по абсолютной величине) коэффициента корреляции между этими переменными [14, с. 155].

ное управление значениями Y путем регулирования величин объясняющих переменных X. Такая постановка задачи претендует на проникновение в «физический механизм» изучаемых статистических связей, т. е. в тот самый механизм преобразования «входных» переменных X и ε в результирующие показатели Y (см. рис. В.1), который в большинстве случаев исследователь, не будучи в состоянии его конструктивно описать, вынужден именовать (следуя сложившейся кибернетической терминологии) «черным ящиком».

И при выявлении причинных связей, и при намерении исследователя использовать модели типа (В.3) или (В.4) для управления значениями результирующих показателей $Y_{\text{ср}}(X)$ или $Y(X)$ путем регулирования величин объясняющих переменных X на первый план выходит *задача правильного определения структуры модели* (т. е. выбора общего вида функции $\mathbf{f}(X)$), решение которой обеспечивает возможность количественного измерения эффекта воздействия на $Y(X)$ каждой из объясняющих переменных $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$ *в отдельности*. Однако как раз это место (правильный выбор общего вида функции $\mathbf{f}(X)$) и является самым слабым во всей технике статистического исследования зависимостей: к сожалению, не существует стандартных приемов и методов, которые образовывали бы строгую теоретическую базу для решения этой важнейшей задачи (некоторые рекомендации по проведению этого этапа исследования содержатся в гл. 6).

Заметим, что исследователи, пожалуй, чаще других ставят перед собой именно цели типа 3. И в таких прикладных задачах, как *управление качеством продукции* с помощью регулирования хода технологических процессов [95, 47], *прогноз и анализ объемов произведенной продукции* по затратам на трудовые ресурсы и капитальные вложения [31, 152], *построение интегральных целевых функций*, описывающих эффективность функционирования экономических единиц (предприятий, семей) по набору частных характеристик [9, 11, 128] и др., это вполне оправдано. Однако, к сожалению, далеко не всегда целевые установки исследователей подкреплены объективными возможностями их реализации.

В.3. Математический инструментарий

Методы статистического исследования зависимостей составляют содержание отдельных частей многомерного статистического анализа, который можно определить [8, с. 731] как раз-

дел математической статистики, посвященный построению оптимальных планов сбора, систематизации и обработки многомерных статистических данных типа (В.1), нацеленных в первую очередь на выявление характера и структуры взаимосвязей между компонентами исследуемого многомерного признака (X, Y) и предназначенных для получения научных и практических выводов. При этом среди $p + m$ компонент исследуемого многомерного признака (X, Y) могут быть: *количественные*, т. е. скалярно измеряющие в определенной шкале степень проявления изучаемого свойства объекта (денежный доход и сбережения семьи, объем валовой продукции, численность работников на предприятии и т. п.); *порядковые* (или *ординальные*), т. е. позволяющие упорядочивать анализируемые объекты по степени проявления в них изучаемого свойства (уровень жилищных условий семьи, квалификационный разряд рабочего, уровень образования работника и т. п.); *классификационные* (или *номинальные*), т. е. позволяющие разбивать обследованную совокупность объектов на не поддающиеся упорядочиванию однородные (по анализируемому свойству) классы (профессия работника, мотив миграции семьи, отрасль промышленности и т. п.). Разделы многомерного статистического анализа, составляющие математический аппарат статистического исследования зависимостей, формировались и развивались с учетом специфики анализируемых моделей, обусловленной природой изучаемых переменных. Соответствующая специализация этих разделов отражена в табл. В.2. В ней же указаны главы данной книги и другие литературные источники, посвященные описанию указанных разделов.

Из табл. В.2 видно, что данная книга не охватывает методов исследования зависимостей неколичественного или смешанного (разнотипного) результирующего показателя от количественных или смешанных объясняющих переменных: объемность и специфичность указанной темы обусловливают целесообразность посвящения ей специального издания.

Кроме того, принцип систематизации различных схем, принятый в табл. В.2, не приспособлен для выделения одного важного (особенно в области социально-экономических приложений) случая, когда связи между количественными переменными X и Y описываются *системой одновременных уравнений*, в которых одни и те же переменные могут играть одновременно (в различных уравнениях системы) и роль результирующих, и роль объясняющих. Этому посвящена *теория одновременных эконометрических уравнений*, основные результаты которой представлены в гл. 14.

Таблица В.2

№ п/п	Природа результирующих показателей	Природа объясняющих переменных (предикторов)	Название обслуживающих разделов многомерного статистического анализа	Главы книги, посвященные данным разделам	Другая литература, посвященная данным разделам
1	Количественная	Количественная	Регрессионный и корреляционный анализ	1, 4, 5, 6, 7, 8, 9, 10, 11, 14	[10, 17, 20, 25, 34, 43, 44, 46, 47, 50, 65; 77, 93, 103, 106, 119]
2	Количественная	Единственная количественная переменная, интерпретируемая как «время»	Анализ временных рядов	12	[18, 21, 28, 41, 66, 80, 144]
3	Количественная	Неколичественная (ординальные или номинальные переменные)	Дисперсионный анализ	13	[66, 148]
4	Количественная	Смешанная (количественные и неколичественные переменные)	Ковариационный анализ, модели типологической регрессии	13	[4, 6, 19, 82]
5	Неколичественная (порядковые, или ординальные, переменные)	Неколичественная (ординальные и номинальные переменные)	Анализ ранговых корреляций и таблиц сопряженности	2,3	[23, 65, 67]
6	Неколичественная (классификационные, или номинальные, переменные)	Количественная	Дискриминантный анализ, кластер-анализ, таксономия, расщепление смесей распределений	—	[11, 19, 20, 48, 58, 66]
7	Смешанная (количественные и неколичественные переменные)	Смешанная (количественные и неколичественные переменные)	Аппарат логических решающих функций	—	[76]

В.4. Некоторые типовые задачи практики

Накопленный опыт практического использования аппарата статистического исследования зависимостей позволяет выделить те типы основных прикладных направлений исследований, в которых этот аппарат работает особенно часто и плодотворно. Если попытаться расщепить общую проблему оптимального управления сложной системой (т. е. центральную проблему кибернетики) на основные составляющие (рис. В.3), то

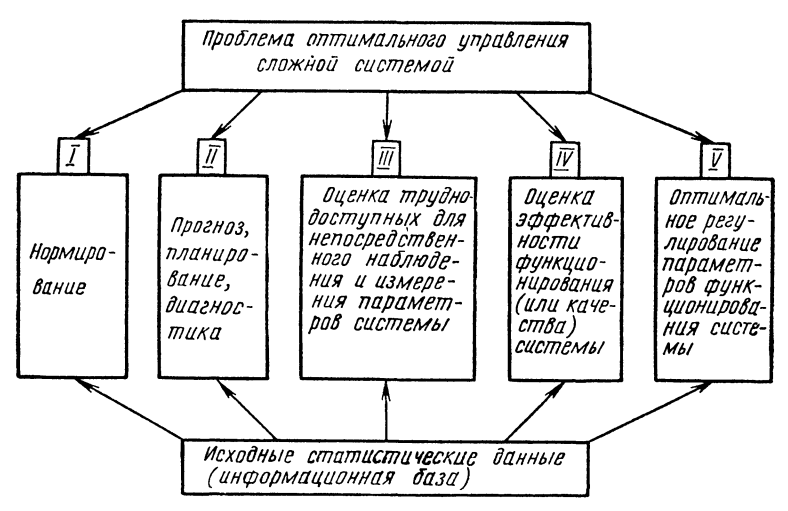


Рис. В.3. Основные направления практического использования аппарата статистического исследования зависимостей и центральная проблема кибернетики

в качестве этих составляющих как раз и фигурируют именно те направления прикладных исследований, в разработке которых существенную роль играет математический аппарат статистического исследования зависимостей.

Естественность предложенного здесь расщепления общей проблемы оптимального управления сложной системой легко пояснить практически на любом примере принятия управленческого решения. Остановимся, скажем, на примере принятия управленческого решения руководителем производственного или учрежденческого подразделения при зачислении в штат нового сотрудника. Основываясь на знании необходимой информационной базы (в данном случае это целевые установки и возможности подразделения и основные сведения о принимаемом сотруднике), лицо, принимающее решение (ЛПР),

должно последовательно проанализировать и решить следующие задачи:

а) определить нормативные требования к деятельности сотрудников, т. е. пронормировать их труд (направление I на рис. В.3);

б) спрогнозировать возможности сотрудника и, сопоставив их с основными целевыми установками подразделения, спланировать его деятельность, включив ее в план общего фронта работ, выполняемых подразделением (направление II);

в) при прогнозировании потенциальных возможностей нового сотрудника (а в ряде случаев — и при последующей оценке эффективности его деятельности) весьма существенным оказывается умение оценить ряд таких не поддающихся непосредственному измерению его качеств, как инициативность, творческая активность, дисциплинированность, трудолюбие, обязательность, «контактность» с другими членами коллектива и т. п. (направление III);

г) в некоторых (особенно непроизводственных) областях деятельности человека оценка эффективности его работы (без которой невозможно оптимальное управление) сводится к весьма трудной задаче построения агрегированного показателя (латентного, т. е. скрытого, непосредственно не измеряемого) ее качества (направление IV);

д) и наконец, опираясь на решение задач а)—г) и на возможность регулирования параметров (в данном случае стимулирующего и «штрафного» характера), от которых в определенной мере и в соответствии с некоторой, как правило, статистической закономерностью зависит уровень эффективности работы сотрудника, ЛПР осуществляет такую «настройку» значений этих параметров, которая обеспечивает, по возможности, оптимальный режим функционирования всей системы, т. е. вверенного ему подразделения (направление V).

Остановимся кратко на роли методов статистического исследования зависимостей в разработке каждого из упомянутых направлений.

I. Нормирование. Общая схема формирования нормативов с использованием методов статистического исследования зависимостей может быть представлена следующим образом. Нормативный показатель играет в моделях типа (В.3)—(В.4) роль результирующей (объясняемой) переменной y, а факторы, участвующие в расчете нормативного показателя, — роль объясняющих (предикторных) переменных $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$. Предполагается, что привлечение для расчета норматива y полной системы определяющих его факторов, т. е. такой си-

стемы, с помощью которой возможно детерминированное (однозначное) определение величины y, либо принципиально невозможно, либо нецелесообразно из-за чрезмерного усложнения расчетных формул. Поэтому анализируется связь между y и $(x^{(1)}, x^{(2)}, \ldots, x^{(p)})$ вида

$$y = f(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; \Theta) + \varepsilon, \tag{B.11}$$

где ε — остаточная случайная компонента, обусловливающая возможную погрешность в определении норматива y по известным значениям факторов $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, а $f(X; \Theta)$ — функция из некоторого известного параметрического семейства $\mathbf{F} = \{f(X; \Theta)\}$, $\Theta \in A$, однако численное значение входящего в ее уравнение параметра Θ (вообще говоря, векторного) неизвестно. С целью подбора «подходящего» значения Θ проводится контрольный эксперимент (наблюдение), в результате которого исследователь получает исходные статистические данные вида (В.1). Далее на основании этих данных проводится необходимый статистический анализ модели (В.11) с целью получения оценки $\widehat{\Theta}$ неизвестного параметра Θ и анализа точности полученной расчетной формулы $\widehat{Y}_{\text{ср}}(X) = f(X; \widehat{\Theta})$, в которой величина условной (экспериментальной) средней $\widehat{Y}_{\text{ср}}(X)$ интерпретируется как средний нормативный показатель при значениях определяющих факторов, равных X.

Данный подход использовался, в частности, при разработке методик расчета численности служащих (по различным их функциям) на промышленном предприятии отрасли по набору технико-экономических показателей, характеризующих предприятие, при построении автоматизированных систем нормирования ремонтных работ [82] и в других областях (см., например, ГОСТ 22015—76 «Качество продукции. Нормирование и статистическая оценка качества металлических материалов и изделий по механическим характеристикам»).

II. Прогноз, планирование, диагностика. Отправляясь от общей формулировки задачи статистического исследования зависимостей (см. § В.1) и от ее модельной записи (В.11), определим в качестве результирующей переменной y интересующий нас прогнозируемый (планируемый, диагностируемый) показатель, а в качестве объясняющих (предикторных) переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ — сопутствующие факторы, значения которых содержат основную информацию о величине этого показателя[1]. Наличие остаточной случайной компоненты

[1] В моделях прогноза и планирования в качестве одного из объясняющих факторов $x^{(k)}$ вводится в явном виде «длина прогноза», или «горизонт планирования», t (в единицах времени).

ε, как и прежде, отражает тот факт, что переменные $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$ содержат не всю информацию об y, и обусловливает неизбежность погрешности в определении прогнозируемого (планируемого, диагностируемого) показателя по известным значениям объясняющих факторов $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$. Исходные статистические данные вида (В.1) исследователь получает, регистрируя одновременно значения y и $(x^{(1)}, ..., x^{(p)})$ на анализируемых объектах в прошлом (в базовом периоде) или на других объектах, но однородных с анализируемыми.

Имеется обширная литература по решению задач прогноза, планирования и диагностики с использованием аппарата статистического исследования зависимостей [4, 29, 31, 47, 80, 93, 128, 144, 152, 163]. В табл. В.3 приведены примеры некоторых типичных задач этого направления прикладных исследований.

Можно было бы продолжить перечень примеров табл. В.3, заполнив их аналогичными задачами из энергетики (задача оперативного и долгосрочного прогноза потребления электроэнергии), гидрологии, социологии, физики и других областей деятельности человека.

III. Оценка труднодоступных для непосредственного наблюдения и измерения параметров системы. Восстановление возраста археологической находки по ряду косвенных признаков; *прочности бетона* с помощью косвенных (неразрушающих) методов контроля (например, по отношению диаметров отпечатков на поверхности испытуемого образца бетона и на воздействующем на него эталонном молотке [16]); *денежных сбережений семьи* по ее доходу (в среднедушевом исчислении) — во всех этих ситуациях исследователь вынужден иметь дело с показателями, труднодоступными для непосредственного измерения (они выделены в тексте курсивом). Очевидно, для того чтобы иметь принципиальную возможность статистически выявить связь, существующую между труднодоступным показателем y и косвенно связанными с ним, но легко поддающимися наблюдению и измерению признаками $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$, исследователю необходимо располагать исходными статистическими данными вида (В.1), которые получают с помощью специально организованного контрольного эксперимента или наблюдения [16]. После того как эта связь выявлена (и оценена степень ее точности), она используется для косвенного определения значений труднодоступных показателей лишь по значениям объясняющих переменных $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$.

IV. Оценка эффективности функционирования (или качества) анализируемой системы. Пытаясь оценить (в целом) эффективность деятельности отдельного специалиста, подраз-

деления или предприятия, проранжировать страны по некоторому интегральному качеству (например, по степени прогрессивности структуры их фондов потребления или всего национального дохода [11]), наконец, проставить балльные оценки спортсмену — участнику командных соревнований в игровых видах спорта за качество его игры в определенном цикле [11], мы каждый раз, по существу, решаем (на интуитивном уровне) одну и ту же задачу: отправляясь в своем анализе от набора частных показателей $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, каждый из которых может быть измерен и характеризует какую-нибудь одну частную сторону понятия «эффективность», мы их как бы взвешиваем (т. е. внутренне оцениваем удельный вес их влияния на общее, агрегированное, понятие эффективности) и выходим на некоторый скалярный агрегированный показатель эффективности y. Этот показатель — латентный (скрытый), так как он принципиально не поддается непосредственному измерению (не существует или нам не известна объективная шкала, в которой он мог бы быть измерен). Но он с некоторой точностью восстанавливается по значениям частных показателей эффективности $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$. Это значит, что между латентным агрегированным показателем y и набором частных критериев эффективности $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ существует статистическая связь типа (В.11).

Главная особенность (и трудность) описываемой ситуации заключается в том, что при получении (сборе) исходной статистической информации вида (В.1) значения результирующего показателя y могут быть получены только с помощью специально организованного экспертного опроса (значения частных критериев эффективности $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, как правило, поддаются непосредственному измерению). Форма экспертной информации о значениях y может быть различной (балльные оценки, упорядочения, парные сравнения [11]). Но только располагая наряду со статистической информацией об $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})'$ одной из форм соответствующей экспертной информации об y, мы можем статистически построить некоторую аппроксимацию $\widehat{y}_{\text{ср}}(X) = f(X; \widehat{\Theta})$ для агрегированного критерия эффективности функционирования системы и использовать ее затем в качестве формализованного метода оценки интегрального понятия эффективности (т. е. уже без привлечения экспертов, а лишь по частным критериям $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$). Такая модифицированная форма использования аппарата статистического исследования зависимостей предложена в [9], развита в [68] и носит название *экспертно-статистического метода построения неизвестной целевой функции*.

Таблица В.3

№ п/п	Содержание задачи	Прогнозируемый (планируемый, диагностируемый) показатель y	Предсказывающие (объясняющие) переменные $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$	Аналитическая запись общего вида исследуемой зависимости (один из вариантов)	Литература, посвященная данной задаче
1	2	3	4	5	6
1	Прогноз и планирование объема выпускаемой продукции по факторам производства (построение производственных функций)	Объем валовой продукции	$x^{(1)}$ — затраты на труд; $x^{(2)}$ — затраты на капитальные вложения; $x^{(3)}$ — время (номер года)	$y_{ср}(X)=\theta_0 (x^{(1)})^{\theta_1} \times (x^{(2)})^{\theta_2} e^{\theta_3 x^{(3)}}$	[31, 47, 80 126, 152]
2	Прогноз урожайности сельскохозяйственных культур по климатическим факторам и факторам сельскохозяйственного производства	Урожайность	$x^{(1)}$ — сумма весенних «активных температур»; $x^{(2)}$ — количество весенних осадков; $x^{(3)}$ — механовооруженность; $x^{(4)}$ — затраты на удобрения	$y_{ср}(X)=\theta_0 \cdot \prod_{k=1}^{4} (x^{(k)})^{\theta_k}$ или $y_{ср}(X)=\theta_0+\sum_{k=1}^{4} \theta_k x^{(k)}$	[65]
3	Прогноз производительности труда, анализ ее динамики	Производительность труда	$x^{(1)}$ — фондовооруженность; $x^{(2)}$ — энерговооруженность; $x^{(3)}$ — время (могут привлекаться и другие факторы с учетом специализации производства)	Та же, что и в п. 1	[144]

4	Прогноз объемов потребления продукции или услуг определенного вида (построение кривых Энгеля)	Удельная величина спроса (потребления) товаров или услуг определенного вида	x — среднедушевой доход	$y_{ср} = \frac{\theta_0}{1 + \theta_1 e^{-\theta_2 x}}$ — логистическая кривая ($\theta_2 > 0$)	[128]
5	Анализ динамики национального дохода и взаимосвязей его основных составных частей	$y_t^{(1)}$ — доход в году t; $y_t^{(2)}$ — фонд потребления в году t	x_t — капиталовложения (инвестиции) в году t	$y_t^{(1)} = y_t^{(2)} + x_t$; $y_t^{(2)} = \theta_0 + \theta_1 y_t^{(1)} + \varepsilon_t$ (ε_t — остаточная случайная компонента)	[29, 31, 80]
6	Техническая диагностика	Показатель технического состояния системы или процесса	Значения параметров системы или процесса, косвенно характеризующих различные частные аспекты ее технического состояния	Зависит от специфики задачи	[5, 145]
7	Медицинская диагностика	Наличие («тяжесть») заболевания	Результаты медико-биологических анализов и тестирование пациентов	Зависит от специфики задачи	[163]

Продолжение табл. В.3

№ п/п	Содержание задачи	Прогнозируемый (планируемый, диагностируемый) показатель y	Предсказывающие (объясняющие) переменные $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$	Аналитическая запись общего вида исследуемой зависимости (один из вариантов)	Литература, посвященная данной задаче
1	2	3	4	5	6
8	Геологический прогноз (месторождений)	Наличие (уровень) рудоносности в исследуемом месте	Процентное содержание ряда сопутствующих элементов в исследуемом месте, их динамика в «геологическом» времени	Зависит от специфики задачи	
9	Прогноз и планирование конструкционных и технико-экономических характеристик проектируемого сооружения	Конструкционные и технико-экономические характеристики проектируемого сооружения	Исходные параметры места и условий строительства, нормативные задания по основным результирующим показателям проектируемого сооружения	Зависит от специфики задачи	[4]
10	Прогноз и планирование надежностных характеристик отдельных узлов и элементов сложного изделия	Долговечность (продолжительность жизни до разрушения) элемента	x — величина эксплуатационного напряжения	$y_{ср}(x) = \theta_0 + \theta_1 x^{-\theta_2}$ $(\theta_2 > 0)$	[10, 125]

В описанную схему вкладывается широкий класс задач теории и практики измерения комплексного понятия «качество» сложного изделия (т. е. квалиметрии [5]): в этих задачах y интерпретируется как агрегированный (комплексный) показатель качества изделия, а $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ — как отдельные частные характеристики его качества (надежность, экономичность, удобство пользования, эстетический вид и т. п.). В качестве параметрических семейств $\mathbf{F} = \{f(X; \Theta)\}$, привлекаемых при статистическом анализе задач данного типа, чаще других используются функции *линейные*

$$f(X; \Theta) = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)} \qquad (B.12)$$

и степенные

$$f(X; \Theta) = \theta_0 (x^{(1)})^{\theta_1} (x^{(2)})^{\theta_2} \ldots (x^{(p)})^{\theta_p} \qquad (B.13)$$

последняя особенно характерна для задач квалиметрии).

Остается отметить, что и традиционные подходы аппарата статистического исследования зависимостей (классический регрессионный анализ, метод наименьших квадратов и т. п.) широко используются в практике оценки технического уровня и качества продукции. Это, в частности, отражено и в соответствующей официальной документации (см., например, РД 50—149—79: Методические указания по оценке технического уровня и качества промышленной продукции. Основные положения; ГОСТ 22732—77: Методы оценки уровня качества промышленной продукции и др.).

V. Оптимальное регулирование параметров функционирования анализируемой системы. Рассмотрим пример [10]. При анализе производительности мартеновских печей на одном из заводов исследовалась, в частности, зависимость между производительностью в тонно/часах (для исключения влияния задержек и простоев часовая производительность мартеновской печи определялась как частное от деления массы плавки на продолжительность периода от начала завалки до выпуска) и процентным содержанием углерода в металле по расплавлении ванны (пробу брали через час после первого скачивания шлака). Результаты замеров по 130 плавкам (т. е. объем n обрабатываемой статистической выборки вида (B.1) равен 130) приведены на рис. B.4. Очевидно, величины производительности (y_i) и процентного содержания углерода (x_i) подвержены некоторому неконтролируемому разбросу, обусловленному влиянием множества не поддающихся строгому учету и контролю факторов. Другими словами, последовательность пар чисел (x_i, y_i), $i = 1, 2, \ldots, 130$, представляет в данном случае ре-

зультаты 130 независимых наблюдений двумерной случайной величины (ξ, η). Однако сквозь кажущуюся хаотичность расположения точек (x_i, y_i) на рис. В.4 просматривается вполне определенная закономерность зависимости условного среднего значения производительности $y_{\text{ср}}(x) = \mathbf{E}(\eta | \xi = x)$ от величины процентного содержания углерода x. Поэтому, располагая статистической зависимостью $y_{\text{ср}}(x)$, мы можем дать рекомендации технологу по оптимальному (с точки зрения максимизации производительности) управлению процессом выплавки: поддерживать процентное содержание углерода в пределах 0,6—1,0%.

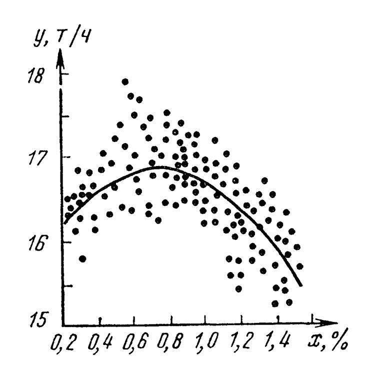


Рис. В.4. Зависимость производительности (y, т/ч) от процентного содержания углерода (х, %) в металле до расплавления

Мы не случайно начали с этого примера. Использование методов статистического исследования зависимостей в задачах оптимального регулирования хода технологического процесса и построения соответствующих автоматизированных систем управления технологическими процессами (АСУТП) можно отнести к примерам грамотных и относительно распространенных актуальных приложений этого аппарата [47, 145]. Общая схема таких приложений предусматривает (в дополнение к приведенному выше частному примеру): а) одновременное рассмотрение нескольких результирующих показателей $y^{(1)}, y^{(2)}, \ldots, y^{(m)}$ (производительность, качество продукции, расход сырья и энергии и т. п.) и многих регулируемых параметров технологического процесса $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$; б) возможность сбора исходной статистической информации вида (В.1) в условиях активного эксперимента (см. § В.1, задача 7).

Менее освоенным (но не менее правомерным и актуальным) является этот подход в задачах оптимального регулирования:

характеристик социально-экономического поведения людей и целых коллективов в ситуациях, когда существует принципиальная возможность выявления статистических связей между этими характеристиками и набором объясняющих (и хотя бы частично регулируемых) факторов [40, 128];

характеристик курса медицинского лечения;

структуры и объемов нагрузок и видов заданий в процессе профессиональной подготовки специалистов.

В.5. Основные типы зависимостей между количественными переменными

При изучении взаимосвязей между анализируемыми количественными показателями следует установить, к какому именно типу зависимостей относится исследуемая схема. Под типом зависимости мы подразумеваем в данном случае не аналитический вид функции $Y_{\text{ср}}(X) = f(X; \Theta)$ в моделях вида (В.11) (о выборе общего аналитического вида функции $f(X; \Theta)$ см. гл. 6), а природу анализируемых переменных (X, y) и соответственно интерпретацию функции $f(X; \Theta)$ в каждом конкретном случае.

Зависимость между неслучайными переменными (схема A). В этом случае результирующий показатель y детерминированно (т. е. вполне определенно, однозначно) восстанавливается по значениям неслучайных объясняющих переменных $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$, т. е. значения y зависят только от соответствующих значений X и полностью ими определяются. Это — обычная схема чисто функциональной зависимости между неслучайными переменными, когда y является некоторой функцией от p переменных X (т. е. $y = f(X)$), что является вырожденным случаем зависимостей вида (В.11), когда остаточная случайная компонента ε равна нулю (с вероятностью единица).

Известно, например, что возраст дерева y (в годах) можно однозначно восстановить по числу колец x на срезе его ствола, а именно $y = x$. Примеры адекватного описания реальных зависимостей с помощью чисто функциональных (нестохастических) связей, к сожалению, крайне редки в практике исследований. Кроме того, при проведении их анализа нет необходимости использовать методы вероятностно-статистической теории. Поэтому в дальнейшем изложении мы не будем больше возвращаться к этому типу зависимостей.

Регрессионная зависимость случайного результирующего показателя η от неслучайных предсказывающих переменных X (схема B). Природа такой связи может носить двойственный характер: а) регистрация результирующего показателя η неизбежно связана с некоторыми случайными ошибками измерения ε, в то время как предикторные (объясняющие) переменные $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})'$ измеряются без ошибок (или величины этих ошибок пренебрежимо малы по сравнению с соответствующими ошибками измерения результирующего показателя); б) значения результирующего показателя η зависят не только от соответствующих значений X, но и еще от

ряда неконтролируемых факторов, поэтому при каждом фиксированном значении X^* соответствующие значения результирующего показателя $\eta(X^*) = (\eta | X = X^*)$ неизбежно подвержены некоторому случайному разбросу.

В этом случае предикторные переменные X играют роль неслучайного (векторного при $p > 1$) параметра, от которого зависит закон распределения вероятностей (в частности, среднее значение и дисперсия) исследуемого результирующего показателя η. Удобной математической моделью такого рода зависимостей является разложение вида

$$\eta(X) = f(X) + \varepsilon(X), \tag{B.14}$$

в котором неслучайная составляющая правой части (функция $f(X)$) описывает поведение условного среднего $y_{\text{ср}}(X) = \mathbf{E}\eta(X) = f(X)$ в зависимости от X, а остаточная случайная компонента $\varepsilon(X)$ отражает случайную природу $\eta(X)$. В широком классе исследуемых схем модель (B.14) строится таким образом, что математическое ожидание случайного остатка $\varepsilon(X)$ равно нулю ($\mathbf{E}\varepsilon(X) \equiv 0$) тождественно по X; предполагается обычно, что при всех X существует конечная дисперсия $\varepsilon(X)$ (т. е. $\mathbf{D}\varepsilon(X) < \infty$), причем величина этой дисперсии, вообще говоря, может зависеть от X (т. е. $\mathbf{D}\varepsilon(X) = \sigma^2(X)$). Подчеркнем то обстоятельство, что в описанной модели (B.14) ни природа случайной компоненты $\varepsilon(X)$, ни соответственно характеристики ее вероятностного распределения никак не связаны со структурой функции $f(X)$ и, в частности, не зависят от значений ее параметра Θ в параметрической записи модели (т. е. когда вместо всех возможных функций $f(X)$ рассматривают какое-либо параметрическое семейство $f(X; \Theta)$, см., например, (B.12), (B.13)).

Если вернуться к примеру B.1, то можно убедиться, что он хорошо укладывается в рамки модели (B.14). Для этого следует лишь заметить, что имевшаяся в этом примере возможность контролировать значения предикторной переменной ξ, по существу, переводит эту переменную из категории случайных величин в категорию неслучайных (контролируемых) параметров модели. Дальнейший анализ примера B.1 (см. табл. B.1, формулу (B.5) и рис. B.2) подсказал нам следующую конкретизацию допущений о природе составных частей модели (B.14):

$$\begin{aligned} y_{\text{ср}}(x) &= \mathbf{E}\eta(x) = f(x) = \theta_0 + \theta_1 x; \\ \sigma^2(x) &= \mathbf{D}\varepsilon(x) = \sigma_0^2 \cdot (y_{\text{ср}}(x))^2, \end{aligned} \tag{B.15}$$

где σ_0^2 — константа, не зависящая от x.

Пример В.2. В табл. В.4 и на рис. В.5 представлены результаты усталостных испытаний алюминиевых сплавов [125], т. е. набор сорока пар (x_i, y_i), $i = 1, 2, \ldots, 40$, экспериментальных значений величин x и η соответственно.

Если при сборе выборочных данных, составляющих двумерную систему наблюдений, производится по нескольку наблюдений при каждом фиксированном значении аргумента, а также в случае разбиения диапазона переменной — аргумента на интервалы группирования $\Delta_i^{(x)}$, в общую схему обозначений двумерной системы наблюдений (В.1) целесообразно внести некоторые изменения.

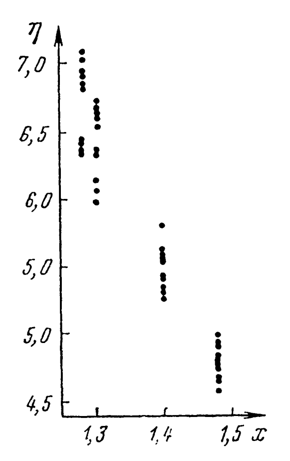


Рис. В.5. Графическое представление результатов усталостных испытаний алюминиевых сплавов

Так, если k — число различных фиксированных значений предикторной переменной (или количество интервалов группирования $\Delta_i^{(x)}$, на которые разбит весь обследованный диапазон этой переменной), а m_i $(i = 1, 2, \ldots, k)$ — количество наблюдений, произведенных при i-м фиксированном значении аргумента (или количество наблюдений, попавших в i-й интервал разбиения $\Delta_i^{(x)}$), то результаты наблюдений удобнее снабдить двумя индексами, т. е. записать в виде (x_{ij}, y_{ij}), где $i = 1, 2, \ldots, k$, а $j = 1, 2, \ldots, m_i$. Здесь первый индекс (i) обозначает порядковый номер фиксированного значения независимой переменной (или порядковый номер интервала группирования), а второй индекс (j) — порядковый номер наблюдения, произведенного при данном i-м фиксированном значении аргумента (или порядковый номер наблюдения, попадающего в i-й интервал группирования). Так, например, под (x_{35}, y_{35}) понимается результат пятого по порядку наблюдения, произведенного при третьем фиксированном значении аргумента (или попадающего в третий интервал группирования $\Delta_3^{(x)}$). В наших рассмотрениях будут фигурировать также величины $x_1^0, x_2^0, \ldots, x_k^0$, представляющие собой последовательность *различных* фиксированных значений аргумента, при которых производились наблюдения (или средние точки интервалов группирования $\Delta_i^{(x)}$), а также условные средние зависимой переменной

Таблица В.4

i	x_i^0	m_i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}
1	1,28	10	6,34	6,34	6,41	6,42	6,80
2	1,30	10	5,95	6,04	6,11	6,31	6,36
3	1,40	10	5,23	5,27	5,32	5,39	5,40
4	1,48	10	4,55	4,65	4,65	4,68	4,72

i	y_{i6}	y_{i7}	y_{i8}	y_{i9}	y_{i10}	$\bar{y}_i$	$s^2(x_i^0)$
1	6,85	6,91	6,91	7,02	7,12	6,71	0,091
2	6,52	6,60	6,62	6,64	6,71	6,39	0,076
3	5,52	5,52	5,53	6,60	5,78	5,46	0,020
4	4,73	4,78	4,78	4,84	4,86	4,72	0,009

$$\bar{y}(x_i^0) = \bar{y}_i = \frac{1}{m_i}\left(y_{i1} + y_{i2} + \dots + y_{im_i}\right),$$

характеризующие средние значения результирующего показателя при каждом фиксированном значении аргумента x_i^0 (или средние значения в каждом отдельном интервале группирования $\Delta_i^{(x)}$).

Очевидно, что в ситуациях, когда производится по нескольку наблюдений при каждом фиксированном значении аргумента, мы будем иметь

$$x_{11} = x_{12} = \dots = x_{1m_1} = x_1^0;$$

$$x_{21} = x_{22} = \dots = x_{2m_2} = x_2^0 \text{ и т. д.}$$

В качестве результирующего показателя — случайной переменной η в нашем примере рассматривается характеристика долговечности образца — нормированная величина логарифма числа циклов N до разрушения образца, а в качестве неслучайной предикторной переменной x — логарифм соответствующей величины эксплуатационного напряжения V, Н/мм² (кг/мм²). Очевидно, долговечность образца зависит также от целого ряда неконтролируемых факторов (случайное варьирование условий эксперимента, свойств самих образцов и т. п.), поэтому при каждом уровне напряжения характеристики долговечности будут подвержены некоторому случайному разбросу около своего среднего.

Расположение экспериментальных точек (x_i, y_i) на рис. В.5 указывает на систематическую закономерность в поведении условных средних $\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$ в зависимости от номера i, т. е. от величины x; их расположение близко к прямолинейному. Это приводит к гипотезе о целесообразности представления исследуемой случайной величины выражением (В.15). Первыми шагами исследователя может быть приближенная оценка прямой $y_{\text{ср}}(x) = \theta_0 + \theta_1 x$, а также меры случайного разброса индивидуальных значений η вокруг этой прямой, характеризующейся в первом приближении только эмпирическими дисперсиями $s^2(x_i^0)$. Однако при проведении более точного количественного анализа возникают следующие вопросы: как наиболее точно провести прямую $y_{\text{ср}}(x) = \theta_0 + \theta_1 x$; как оценить степень точности построенной зависимости; нельзя ли строить математически обоснованные зоны (так называемые доверительные интервалы и границы) около исследуемой прямой, попадание в которые эмпирических индивидуальных или средних значений η при каждом фиксированном x гарантировалось бы с заранее заданной вероятностью? Ответы на все эти вопросы и дает регрессионный анализ (см. гл. 5—11).

Корреляционно-регрессионная зависимость между случайными векторами η — результирующим показателем и ξ — предикторной переменной (схема C). В данном типе моделей и компоненты вектора результирующего показателя η, и компоненты вектора объясняющих переменных ξ зависят от множества неконтролируемых факторов, так что являются случайными по своей физической сущности. Мы уже сталкивались с такой ситуацией в примере, в котором исследовалась связь между производительностью мартеновских печей и процентным содержанием углерода в металле (см. рис. В.4). Зависимости такого типа вообще характерны для описания хода технологических процессов, реальные значения параметров которых $\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})'$, равно как и характеризующие их результирующие показатели $\eta = (\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(m)})'$, как правило, флюктуируют случайным (но взаимосвязанным) образом около установленных номиналов.

В подобных ситуациях оказывается полезным рассмотреть разложение исследуемого результирующего показателя η на две случайные составляющие по формуле типа (В.3). Первая из них определяется некоторой (векторнозначной) функцией $\mathbf{f}$ от объясняющей переменной ξ, а вторая отражает остаточные влияния неучтенных случайных факторов на анализируемый

результирующий показатель η. Итак

$$\eta = \mathbf{f}(\xi) + \varepsilon. \quad (В.16)$$

При этом разложение (В.16) строится таким образом, чтобы для компонент векторов $\mathbf{f}(\xi)$ и ε выполнялись соотношения $\mathbf{E}\varepsilon^{(k)} = 0$, $\mathbf{D}\varepsilon^{(k)} = \sigma_k^2 < \infty$, $\operatorname{cov}(f^{(k)}(\xi), \varepsilon^{(k)}) = \mathbf{E}[(f^{(k)}(\xi)\cdot\varepsilon^{(k)}] - \mathbf{E}f^{(k)}(\xi)\cdot\mathbf{E}\,\varepsilon^{(k)} = 0$.

В частном случае единственного результирующего показателя ($m = 1$) и линейного вида функции $f(\xi)$ имеем:

$$\eta = \theta_0 + \sum_{k=1}^{p} \theta_k \xi^{(k)} + \varepsilon. \quad (В.17)$$

Подразумевая, как и прежде, под $y_{ср}(X) = \mathbf{E}(\eta \mid \xi = X)$ условное математическое ожидание результирующего показателя η (при условии, что объясняющая переменная ξ приняла значение, равное X), мы от (В.17) приходим к линейному уравнению регрессии

$$y_{ср}(X) = \theta_0 + \sum_{k=1}^{p} \theta_k x^{(k)}. \quad (В.18)$$

Возможны случаи, когда вторая (остаточная) компонента в разложении (В.16) с полной мерой достоверности (т. е. с вероятностью единица) равна нулю. При этом исследуемые случайные величины η и ξ оказываются связанными *чисто функциональной зависимостью* $\eta = f(\xi)$, но ее следует отличать от функциональной зависимости неслучайных переменных (см. выше, схема A).

Пример В.3. Рис. В.6 иллюстрирует связь между вакуумом в печи для отжига стекла ξ и процентом брака η в стекольном производстве [10].

Случайные изменения свойств сырья, а также ряда неконтролируемых факторов приводят к случайным колебаниям обеих исследуемых переменных. Однако расположение точек на рис. В.6 свидетельствует о том, что эти колебания взаимосвязаны, подчинены вполне определенной закономерности: «облако» рассеяния вытянуто вдоль некоторой прямой, не параллельной ни одной из координатных осей. Все это подтверждает целесообразность разложения случайной величины η по формуле (В.16) и исследования связи между η и ξ, которая в этом случае носит название корреляционной. К перечисленным вопросам регрессионного анализа (построение конкретного вида зависимости между переменными, различные оценки ее точности) в этом случае присоединяется круг вопросов, связанных

с исследованием степени тесноты связи между этими переменными. Совокупность методов, позволяющих решать эти вопросы, принято называть корреляционным анализом (см. гл. 1—3).

Зависимости структурного типа, или зависимости по схеме конфлюэнтного анализа (схемы D_1 и D_2). В обеих описываемых ниже схемах речь идет о восстановлении искомых зависимостей по искаженным наблюдениям анализируемых переменных, причем, в отличие от регрессионной схемы B, искаженными оказываются при наблюдении не только значения результирующего показателя, но и значения *объясняющих* (*предикторных*) *переменных* $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$. В зависимости от того, между какими именно переменными — неслучайными или случайными — исследуются связи, мы будем иметь соответственно тип связи по схеме D_1 или D_2. Оба эти типа связей упоминаются в специальной литературе как *структурные зависимости* [65, с. 500—557] или как зависимости по схеме *конфлюэнтного анализа* [7, 10]. Таким образом, конфлюэнтный анализ предоставляет исследователю совокупность методов математико-статистической обработки данных, относящихся к анализу априори постулируемых функциональных связей между количественными (случайными или неслучайными) переменными $Y = (y^{(1)}, \ldots, y^{(m)})'$ и $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})'$ в условиях, когда наблюдаются не сами переменные, а случайные величины

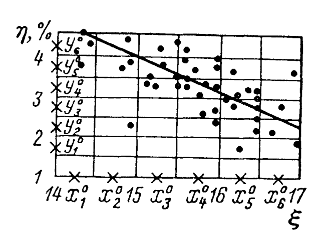


Рис. В.6. Графическое представление данных по связи вакуума в печи для обжига стекла (ξ) и процента брака в стекольном производстве (η)

$$\xi_i^{(k)} = x_i^{(k)} + \varepsilon_{x_i}^{(k)}, \quad k = 1, 2, \ldots, p;$$

$$\eta_i^{(j)} = y_i^{(j)} + \varepsilon_{y_i}^{(j)}, \quad j = 1, 2, \ldots, m; \quad i = 1, 2, \ldots, n, \tag{B.19}$$

где $\varepsilon_{x_i}^{(k)}$ и $\varepsilon_{y_i}^{(j)}$ — случайные ошибки измерений соответственно переменных $x^{(k)}$ и $y^{(j)}$ в i-м наблюдении, а n — общее число наблюдений. При этом общий вид исследуемых функциональных (структурных) связей

$$\begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{pmatrix} = \begin{pmatrix} f^{(1)}(x^{(1)}, \ldots, x^{(p)}; & \Theta) \\ . \quad . \quad . \quad . \quad . \quad . \quad . & . \\ f^{(m)}(x^{(1)}, \ldots, x^{(p)}; & \Theta) \end{pmatrix} \tag{B.20}$$

между ненаблюдаемыми, а точнее, наблюдаемыми с ошибками переменными считается заданным (неизвестным является лишь значение векторного параметра $\Theta = (\theta_1, \ldots, \theta_N)$, участвующего в уравнениях искомых зависимостей (В.20)).

Схема D_1: исследуемые переменные $X = (x^{(1)}, \ldots, x^{(p)})'$ и $Y = (y^{(1)}, \ldots, y^{(m)})'$ не случайны. Для упрощения обозначений проанализируем зависимости (В.19)—(В.20) в рамках данной схемы лишь для *одного* результирующего показателя и *одной* объясняющей переменной (случай $m = 1$, $p = 1$): обобщение этого анализа на случай $m > 1$ и $p > 1$ не представляет принципиальных трудностей.

Учитывая формулы (В.19) и (В.20) и воспользовавшись формальным разложением функции $f(\xi - \varepsilon_x)$ в ряд Тейлора около точки ξ, получаем соотношение между η и ξ:

$$\eta = f(\xi) + \left(\varepsilon_y + \sum_{k=1}^{\infty} (-1)^k \frac{(\varepsilon_x)^k}{k!} f^{[k]}(\xi)\right). \tag{В.21}$$

Здесь под $f^{[k]}(\xi)$ подразумевается k-я производная функции $f(t)$ по t, взятая в точке $t = \xi$. В частности, при линейном виде имеем

$$\eta = (\theta_0 + \theta_1 \xi) + (\varepsilon_y - \theta_1 \varepsilon_x). \tag{В.22}$$

Из (В.21) непосредственно следует, что уравнение регрессии η по ξ (т. е. вид зависимости условного математического ожидания $y_{\text{ср}}(x) = \mathbf{E}(\eta | \xi = x)$ от x) совпадает со структурным соотношением (В.20)[1]. Однако в схеме D_1, в отличие от схем B и C, остаточная случайная компонента в разложениях (В.21) и (В.22) (т. е. соответственно $\varepsilon_y + \sum_{k=1}^{\infty} \frac{(-\varepsilon_x)^k}{k!} f^{[k]}(\xi)$ и $\varepsilon_y - \theta_1 \varepsilon_x$) *зависит от неизвестных параметров*, участвующих в описании функции $f(x)$ и оцениваемых на основании имеющихся у нас выборочных данных.

Эта специфичность природы зависимости, присущая схеме D_1, сильно усложняет задачу построения хороших оценок для неизвестных параметров, входящих в соотношение (В.20). Дело в том, что достаточно хорошо разработанная теория построения таких оценок для схем B и C, в частности оценок максимального правдоподобия, оценок наименьших квадратов,

---

[1]Чтобы убедиться в этом, надо при вычислении условного математического ожидания от обеих частей соотношения (В.21) лишь учесть, что условие $\xi = x$ равносильно условию $\varepsilon_x = 0$, и, кроме того, воспользоваться естественным допущением: $\mathbf{E}\varepsilon_y = \mathbf{E}(\varepsilon_y | \xi = x) = 0$.

оказывается неприменимой к задачам схемы D_1. Так, например, оценки, используемые в регрессионном и корреляционном анализах, при обращении к задачам схемы D_1 теряют свои «хорошие» свойства — несмещенность, эффективность и даже состоятельность. Поэтому исследователь должен проявить особую аккуратность на самой первой стадии анализа — при постановке задачи и определении, к какому из известных типов зависимостей следует отнести данный конкретный случай. Соответственно при описании рекомендаций и приемов

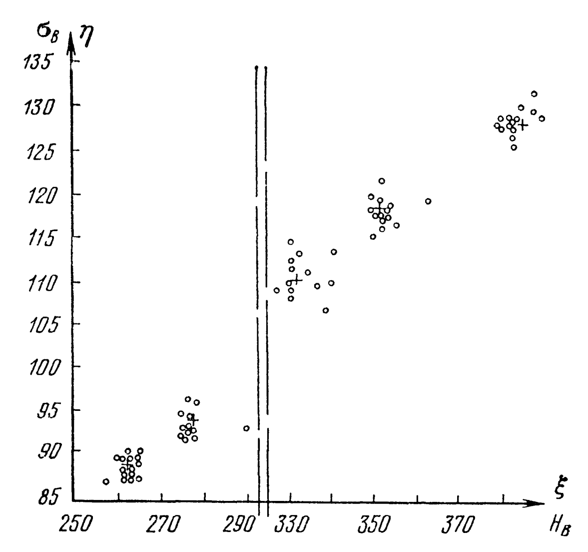


Рис. В.7. Зависимость между пределом прочности σ_b (кг/мм²) и твердостью по Бринелю $H_в$ (кг/мм²) для 75 образцов одной из плавок стали

обработки выборочных данных с целью статистического исследования зависимостей приходится отделять регрессионный и корреляционный анализы (схемы B и C) от так называемого конфлюэнтного анализа (схемы D_1 и D_2).

Пример В.4. [90] На рис. В.7 и в табл. В.5 приведены результаты испытаний образцов (изготовленных из стали 30ХГСА) на твердость по Бринелю (H_B) и предел прочности $(\sigma_b)'$ в Н/мм² (кг/мм²).

Известно, что при существующих условиях производства и конструирования возможность взаимного перевода показателей прочности и твердости для конструкционных сталей (т. е. возможность взаимного сопоставления этих характеристик типа $H_B \rightleftarrows \sigma_b$) зачастую является необходимой. Такой пере-

Таблица В.5

Номер образца	Значения H_B и σ_b при разной термической обработке $(\omega_2^{(i)})$									
	$\omega_2^{(1)}$		$\omega_2^{(2)}$		$\omega_2^{(3)}$		$\omega_2^{(4)}$		$\omega_2^{(5)}$	
	H_B	σ_b	H_B	σ_b	H_B	σ_b	H_B	σ_b	H_B	σ_b
1	263	88,5	277	95,5	331	109,0	363	120,0	383	126,0
2	262	90,0	275	95,5	335	111,5	356	117,0	383	127,0
3	262	90,5	278	94,0	331	109,5	350	118,5	385	128,0
4	262	87,5	278	93,5	331	109,5	352	117,5	390	129,0
5	263	88,5	277	97,0	341	114,0	354	119,0	383	128,0
6	260	90,0	290	93,5	331	113,0	352	122,0	388	132,0
7	263	87,5	277	94,5	331	110,5	350	116,0	383	128,0
8	262	88,0	275	92,5	339	107,5	352	117,0	380	128,0
9	265	90,5	277	93,5	333	114,5	352	118,0	380	128,0
10	262	88,5	277	94,5	331	112,0	350	120,5	385	130,0
11	257	87,5	275	93,5	331	110,0	352	118,5	388	130,0
12	265	90,0	278	96,5	331	115,0	354	118,0	383	128,0
13	265	87,5	278	92,5	337	110,0	350	118,5	380	129,0
14	265	89,5	275	92,5	341	110,5	352	120,0	383	129,0
15	263	90,0	277	93,5	327	109,5	354	119,0	384	128,5
Средние значения $\bar{H}_B$ и $\bar{\sigma}_b$	262,5	88,9	277,6	94,0	333,4	111,1	352,9	118,6	383,9	128,6

вод осуществляется с помощью специальных таблиц, общей основой которых является предположение, что между значениями H_B, σ_b и H_{Rc} (твердость по Роквеллу) существует чисто функциональная взаимно-однозначная зависимость (т. е. зависимость по схеме *А* в нашей классификации). Однако при практическом использовании переводных таблиц и формул было обнаружено, что фактические значения механических характеристик часто существенно отличаются от полученных переводом (даже в тех случаях, когда эти таблицы носят узкоспециализированный характер, т. е. когда они составляются и используются лишь для какого-то одного типа полуфабриката и для одной и той же марки стали).

Причина же подобной рассогласованности, неточности этих таблиц кроется на самом деле в том, что сама природа связи, существующей между различными механическими характеристиками материалов, например между H_B и σ_b, но-

сит не функциональный (детерминированный), а стохастический характер. Так, например, на рис. В.7 видно, что при каждом фиксированном значении твердости соответствующие значения предела прочности σ_b подвержены некоторому неконтролируемому разбросу.

Более детальный профессионально-статистический анализ [90] приводит нас в данном случае к следующей схеме.

На значения H_B и σ_b, так же как и на вид связи, существующей между ними, влияют следующие факторы:

1) химический состав плавки ω_1;

2) термическая обработка ω_2;

3) особенности исследуемого образца — локальный химсостав, размеры зерна в зоне отпечатка, локальная термическая обработка и т. п. ω_3;

4) погрешности измерения, связанные с приборами, установкой образца и т. п. ω_4.

Если величину твердости по Бринелю (H_B) обозначим ξ, а соответствующую величину предела прочности (σ_b) η, то можно воспользоваться выражением, где роль неслучайных (структурных) компонент x и y играют значения ξ и η, взятые для некоторой фиксированной плавки (ω_1) при некотором фиксированном режиме термической обработки (ω_2) и усредненные по всевозможным комбинациям факторов ω_3 и ω_4 (их «наблюденные значения», полученные усреднением по пятнадцати однородным плавкам, изображены на рис. В.7). Что касается остаточных случайных компонент ε_x и ε_y, то наличие каждой из них обусловлено в данном случае различиями в особенностях исследуемых образцов (фактор ω_3). При этом из наших определений следует, что $\mathbf{E}\varepsilon_x = \mathbf{E}\varepsilon_y = 0$. Кроме того, специфика данной конкретной задачи такова, что мы вправе принять в качестве исходных предпосылок для дальнейшего исследования следующие допущения:

а) между структурными компонентами y и x имеется линейная зависимость вида В.15, причем коэффициенты θ_0 и θ_1, вообще говоря, зависят от химического состава (от фактора ω_1), т. е. могут меняться при переходе от одной плавки к другой;

б) пары случайных величин $(\varepsilon_x, \varepsilon_y)$ не зависят друг от друга;

в) при любых фиксированных ω_1 и ω_2 (т. е. для любой фиксированной плавки и при любом фиксированном режиме ее термической обработки) существуют дисперсии $\mathbf{D}\varepsilon_x$ и $\mathbf{D}\varepsilon_y$;

г) «общая» остаточная случайная компонента $\varepsilon = \varepsilon_y - \theta_1\varepsilon_x$ подчинена нормальному распределению, параметры которого не зависят от характера термической обработки (т. е. от фактора ω_2);

д) диапазоны изменения структурных компонент x и y во много раз превосходят практические диапазоны остаточных случайных компонент ε_x и ε_y (см. рис. В.7).

Схема D_2: исследуемые переменные $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})'$ и $\eta = (\eta^{(1)}, \ldots, \eta^{(m)})'$ случайны. Этот тип зависимости, нередко встречающийся в практике статистических исследований, является в некотором смысле обобщением схемы D_1.

Итак, под схемой D_2 мы будем понимать такую схему зависимости, в которой исследуемые случайные переменные ξ и η связаны соотношением (В.20), однако наблюдать мы их можем лишь с некоторыми случайными ошибками — соответственно ε_ξ и ε_η. Поэтому экспериментальными данными (x_i, y_i) в действительности представлены выборочные значения случайных величин (ε', η'), где

$$\xi' = \xi + \varepsilon_\xi, \quad \eta' = \eta + \varepsilon_\eta. \tag{В.23}$$

Обычно предполагают, что ошибки ε_ξ и ε_η взаимно независимы, но зависят от ξ и η и имеют нулевые математические ожидания ($\mathbf{E}\varepsilon_\xi = \mathbf{E}\varepsilon_\eta = 0$) и конечные дисперсии ($\mathbf{D}\varepsilon_\xi < \infty$ $\mathbf{D}\varepsilon_\eta < \infty$).

При этом оказывается, что корреляционные и регрессионные характеристики схемы (ξ', η') могут существенно отличаться от соответствующих характеристик исходной (неискаженной) схемы (ξ, η). Так, например, ниже (см. п. 1.1.4) показано, что наложение случайных нормальных ошибок на исходную двумерную нормальную схему (ξ, η) всегда уменьшает абсолютную величину коэффициента регрессии θ_1 в соотношении (В.15), а также ослабляет степень тесноты связи между ξ и η (т. е. уменьшает абсолютную величину коэффициента корреляции r).

Зависимости по схеме D_2 имеют место, в частности, в задачах исследования хода технологических процессов, когда взаимосвязанные флюктуирующие значения параметров процесса (ξ и η) могут быть измерены лишь с некоторыми случайными ошибками.

В.6. Основные этапы статистического исследования зависимостей

Весь процесс статистического исследования интересующих нас зависимостей удобно разложить на основные этапы. Эти этапы ниже описаны в соответствии с хронологией их реализации, однако некоторые из них находятся, в плане хронологическом, в соотношении итерационного взаимодействия: ре-

зультаты реализации более поздних этапов могут содержать выводы о необходимости повторной «прогонки» (с учетом добытой на предыдущих этапах новой информации) уже пройденных этапов (см., например, схему взаимодействия этапов 3,4, 5 и 6 на рис. В.8). Излагаемая ниже схема приспособлена в основном для исследования зависимостей между количественными переменными, однако с минимальными (и очевидными) модификациями она «работает» и при статистическом анализе связей между неколичественными и разнотипными переменными.

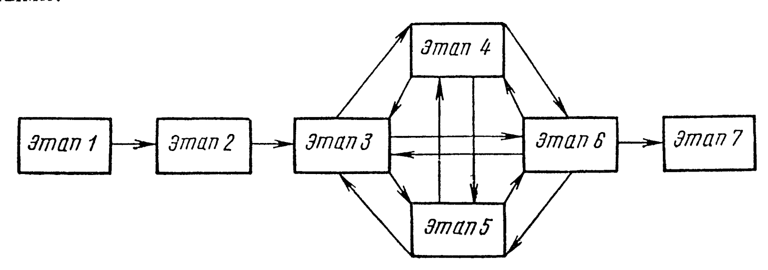


Рис. В.8. Схема хронологически-итерационных взаимосвязей основных этапов статистического исследования зависимостей

Этап 1 (постановочный). Прежде всего исследователь должен определить:

1) элементарную единицу статистического обследования, или элементарный объект исследования O (это может быть страна, город, отрасль, предприятие, семья, индивидуум, пациент, технологический процесс, сложное техническое изделие и т. д.);

2) набор показателей $(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; y^{(1)}, \ldots, y^{(m)})$, регистрируемых на каждом из статистически обследованных объектов, с подразделением их на «входные» (объясняющие) и «выходные» (результирующие) и, если это необходимо, с четким определением способа их измерения; таким образом, на этом этапе каждому элементарному объекту исследования ставится в соответствие перечень анализируемых показателей, т. е.

$$O \longleftrightarrow (x^{(1)}, x^{(2)}, \ldots, x^{(p)}; \quad y^{(1)}, y^{(2)}, \ldots, y^{(m)});$$

3) конечные прикладные цели исследования (см. § В.2), тип исследуемых зависимостей (см. § В.5) и желательную форму статистических выводов (а иногда и степень их точности);

4) совокупность элементарных объектов исследования, на которую мы хотим распространить справедливость действия вы-

явленных в результате анализа статистических зависимостей (если, например, элементарная единица — семья, то анализируемой совокупностью могут быть семьи определенной социальной группы населения или семьи определенной республики и т. д.);

5) общее время и трудозатраты, отведенные на планируемое исследование и коррелированные с ними временная протяженность и объем необходимого статистического обследования (какую часть анализируемой совокупности подвергнуть статистическому обследованию, производить статистическое обследование в статическом или динамическом режиме и т. д.). Заметим, что именно на этом этапе решаются задачи в) и 1, описанные в § В.1.

В решении всех перечисленных вопросов первого этапа исследования главную роль, бесспорно, должен играть «заказчик», т. е. специалист той предметной области, для которой планируется проведение этого исследования.

Этап 2 (*информационный*). Он состоит в проведении сбора необходимой статистической информации вида (В.1). При этом возможны две принципиально различные ситуации: 1) исследователь имеет возможность заранее спланировать выборочное обследование части анализируемой совокупности — выбрать способ отбора элементарных единиц статистического обследования (случайный, пропорциональный, расслоенный и т. д., см., например, [14, п. 5.4.3]), хотя бы по части объясняющих переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ назначить уровни их значений, при которых желательно произвести эксперимент или наблюдения (условия активного эксперимента); 2) исследователь получает исходные данные такими, какими они были собраны без его участия (условия пассивного эксперимента). В любом случае «на выходе» этого этапа исследователь располагает исходными статистическими данными вида (В.1), т. е. каждому (i-му) из статистически обследованных элементарных объектов исследования O_i поставлен в соответствие конкретный вектор характеризующих его «входных» и «выходных» показателей:

$$O_i \longleftrightarrow (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(m)}), \; i = 1,2, \ldots n.$$

(здесь n — общее число статистически обследованных элементарных объектов, т. е. объем выборки). Таким образом, на этом этапе решается, в частности, задача 7 из § В 1.

Говоря о проведении сбора статистических данных, мы не включаем сюда разработку методологии и системы показателей отображаемого объекта: эта работа предполагает про-

фессионально-предметное (экономическое, техническое, медицинское и т. д.) изучение сущности решаемых задач статистического исследования зависимостей, поэтому относится к компетенции соответствующей предметной статистики (экономической и т. д.) и входит в задачи 1-го этапа исследований.

Этап 3 (*корреляционный анализ*). Этот этап нацелен на решение задачи 2 (см. § В.1), он позволяет ответить на вопросы, имеется ли вообще какая-либо связь между исследуемыми переменными, какова структура этих связей и как измерить их тесноту? Описанию методов, с помощью которых проводится такой статистический анализ, посвящены гл. 1—4. Поскольку перечисленные выше вопросы решаются с помощью вычисления и анализа соответствующих корреляционных характеристик, содержание этапа можно определить как проведение корреляционного анализа. Этап достаточно полно оснащён необходимым математическим аппаратом и программным обеспечением, поэтому может быть почти полностью автоматизирован.

Этап 4 (*определение класса допустимых решений*). Главной целью исследователя на этом этапе является определение общего вида, структуры искомой связи между Y и X, или, другими словами, описание класса функций $\mathbf{F}$, в рамках которого он будет производить дальнейший поиск конкретного вида интересующей его зависимости (см. задачи а) и 3 в § В.1). Чаще всего это описание дается в форме некоторого параметрического семейства функций $f(X; \Theta)$, поэтому и этап этот называют также *этапом параметризации модели*. Так, определив в примере В.1, что поиск зависимости среднедушевых семейных сбережений $y_{ср}$ от величины их среднедушевого дохода x мы будем производить в классе $\mathbf{F} = \{\theta_0 + \theta_1 x\}$ линейных функций, мы тем самым завершили четвертый этап исследования (но конкретных числовых значений параметров θ_0 и θ_1 мы к этому моменту еще не знаем).

Следует отметить, что, являясь узловым, в определенной мере *решающим* звеном во всем процессе статистического исследования зависимостей, этот этап в то же время находится в наименее выгодном положении по сравнению с другими этапами (с позиций наличия строгих и законченных математических рекомендаций по его реализации). Поэтому его реализация требует совместной работы специалиста соответствующей предметной области (экономики, техники, медицины и т. д.) и математика-статистика, направленной на как можно более глубокое проникновение в «физический механизм» исследуемой связи. Подходам и методам проведения этого этапа исследований посвящена гл. 6 данного издания.

Существует подход к исследованию моделей регрессии, не требующий предварительного выбора параметрического семейства функций **F** в рамках которого проводится дальнейший анализ. Речь идет о так называемых *непараметрических* (*или частично-параметрических*) методах исследования регрессионных зависимостей, которым посвящена гл. 10. Однако возникающие при их реализации проблемы (необходимость иметь очень большие объемы исходных статистических данных, выбор сглаживающих функций — «окон» и параметров масштаба, выбор порядка сплайна, числа и положения «узлов» и т. п.) сопоставимы по своей сложности с проблемами, возникающими при реализации этапа 4.

Следующие два этапа — 5-й и 6-й — связаны с проведением определенного объема вычислений на ЭВМ и реализуются, по существу, параллельно.

Этап 5 (*анализ мультиколлинеарности предсказывающих переменных и отбор наиболее информативных из них.*) Под явлением мультиколлинеарности в регрессионном анализе понимается наличие тесных статистических связей между предсказывающими переменными $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$, что, в частности, проявляется в близости к нулю (слабой обусловленности) определителя их корреляционной матрицы, т. е. матрицы размера $p \times p$, составленной из парных коэффициентов корреляции $r_{ij} = r(x^{(i)}, x^{(j)})$ ([14, с. 155], а также гл. 1—3 данного издания). Поскольку этот определитель входит в знаменатель выражений для ряда важных характеристик анализируемых моделей (см. гл. 7—11), то мультиколлинеарность создает трудности и неудобства при статистическом исследовании зависимостей по меньшей мере в двух направлениях:

а) в реализации на ЭВМ необходимых вычислительных процедур и, в частности, в крайней неустойчивости получаемых при этом числовых характеристик анализируемых моделей (так, коэффициенты при объясняющих переменных в моделях типа (В.12), (В.13) и др. могут изменяться в несколько раз и даже менять знак при добавлении (или исключении) к массиву исходных статистических данных одного-двух объектов или одной-двух объясняющих переменных);

б) в содержательной интерпретации параметров анализируемой модели, что играет решающую роль в ситуациях, когда конечной целью исследования является цель типа 3 («выявление причинных связей» и т. д., см. § В.2, соотношения (В.9) и (В.9$'$)).

Поэтому исследователь старается перейти к такой новой системе предсказывающих переменных (отобранных из числа исходных переменных $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$ или представленных

в виде некоторых их комбинаций), в которой эффект мультиколлинеарности уже не имел бы места. Этап проводится в основном силами математиков-статистиков с подключением (в самом его конце) специалистов соответствующей предметной области для выбора из нескольких предложенных вариантов набора объясняющих переменных, наиболее легко и естественно интерпретируемого.

Рекомендации по проведению этого этапа даны в гл. 8.

Этап 6 (*вычисление оценок неизвестных параметров, входящих в исследуемое уравнение статистической связи*). Итак, в результате проведения предыдущих этапов были решены, в частности, следующие задачи:

а) определены результирующие и объясняющие переменные и тип исследуемой зависимости (*B*, *C* или *D*, см. § В.5);

б) собрана и подготовлена к счету на ЭВМ исходная статистическая информация вида (В.1);

в) изучены характер и теснота статистических (корреляционных) связей между исследуемыми переменными;

г) выбран класс допустимых решений $\mathbf{F}$, т. е. класс (или параметрическое семейство) функций $\mathbf{f}(X)$, в рамках которого будет подбираться наилучшая (в определенном смысле) аппроксимация $\widehat{\mathbf{f}}(X)$ искомой зависимости типа (В.14), (В.16) или (В.20).

Теперь можно приступать к определению этой наилучшей аппроксимации $\widehat{\mathbf{f}}(X)$, которая является решением оптимизационной задачи вида

$$\widehat{\mathbf{f}}(X) = \arg \operatorname*{extr}_{\mathbf{f} \in \mathbf{F}} \Delta_n(\mathbf{f}), \qquad (В.24)$$

где функционал $\Delta_n(\mathbf{f})$ задает критерий качества аппроксимации результирующего показателя η (или Y) с помощью функции $\mathbf{f}(X)$ из класса $\mathbf{F}$. Выбор конкретного вида этого функционала опирается на знание вероятностной природы остатков ε в моделях типа (В.14), (В.16) и (В.21), причем он строится, как правило, в виде некоторой функции от невязок $\widehat{\varepsilon}_1^{(k)}$, $\widehat{\varepsilon}_2^{(k)}$, ..., $\widehat{\varepsilon}_n^{(k)}$ ($k = 1, 2, \ldots, m$), где $\widehat{\varepsilon}_i^{(k)} = y_i^{(k)} - f^{(k)}(X_i)$ (один из распространенных вариантов такого функционала, а именно *функционал метода наименьших квадратов*, упоминается в примере В.1, см. соотношение (В.7′)). Если в качестве класса $\mathbf{F}$ задаются некоторым параметрическим семейством функций $\{\mathbf{f}(X; \Theta)\}$, то задача (В.24) сводится к подбору (статистическому оцениванию) значений параметров $\widehat{\Theta}$, на которых достигается экстремум по Θ функционала $\Delta_n(\mathbf{f}(X; \Theta))$, а со-

ответствующие модели называют *параметрическими*. Эта часть исследования хорошо оснащена необходимым математическим аппаратом и соответствующим программным обеспечением (см. гл. 7—10).

Этап 7 (*анализ точности полученных уравнений связи*). Исследователь должен отдавать себе отчет в том, что найденная им в соответствии с (В.24) аппроксимация $\hat{\mathbf{f}}(X)$ неизвестной теоретической функции $\mathbf{f}_т(X)$ из соотношений типа (В.14), (В.16) или (В.21) (называемая эмпирической функцией регрессии, см. гл. 5) является лишь некоторым приближением истинной зависимости $\mathbf{f}_т(X)$[1]. При этом погрешность δ в описании неизвестной истинной функции $\mathbf{f}_т(X)$ с помощью $\hat{\mathbf{f}}(X)$ в общем случае состоит из двух составляющих: а) ошибки аппроксимации $\delta_{\mathbf{F}}$ и б) ошибки выборки $\delta(n)$. Величина первой зависит от успеха в реализации этапа 4, т. е. от правильности выбора класса допустимых решений $\mathbf{F}$. В частности, если класс $\mathbf{F}$ выбран таким образом, что включает в себя и неизвестную истинную функцию $\mathbf{f}$ (т. е. $\mathbf{f}_т(X) \in \mathbf{F}$), то ошибка аппроксимации $\delta_{\mathbf{F}} = 0$. Но даже в этом случае остается случайная составляющая (ошибка выборки) $\delta(n)$, обусловленная ограниченностью выборочных данных вида (В.1), на основании которых мы подбираем функцию $\hat{\mathbf{f}}(X)$ (оцениваем ее параметры). Очевидно, уменьшить ошибку выборки мы можем за счет увеличения объема n обрабатываемых выборочных данных, так как при $\mathbf{f}_т(X) \in \mathbf{F}$ (т. е. при $\delta_{\mathbf{F}} = 0$) и правильно выбранных методах статистического оценивания (т. е. при правильном выборе оптимизируемого функционала качества модели $\Delta_n(f)$) ошибка выборки $\delta(n) \to 0$ (по вероятности) при $n \to \infty$ (свойство состоятельности используемой процедуры статистического оценивания неизвестной функции $\mathbf{f}_т(X)$).

Соответственно на данном этапе приходится решать следующие основные задачи анализа точности полученной регрессионной зависимости:

1) в случае $\mathbf{F} = \{f(X;\ \Theta)\}$ и $\mathbf{f}_т(X) \in \mathbf{F}$, т. е. когда класс допустимых решений задается параметрическим семейством функций и включает в себя неизвестную теоретическую функцию регрессии $\mathbf{f}_т(X)$, при заданных доверительной вероятности P и объеме выборки n указать такую предельную

---

[1]В дальнейшем, говоря о вектор-функции $\mathbf{f}_т(X)$, вектор-погрешности δ и векторе результирующих показателей $Y(X)$, мы будем иметь в виду каждую из их компонент в отдельности.

(гарантированную) величину погрешности $\delta_{P,n}(\theta_k)$ для любой компоненты неизвестного векторного параметра Θ, что

$$|\theta_k - \widehat{\theta}_k| \leqslant \delta_{P,n}(\theta_k)$$

с вероятностью, не меньшей, чем P (здесь θ_k — истинное значение k-й компоненты неизвестного параметра Θ, а $\widehat{\theta}_k$ — его статистическая оценка);

2) при заданных доверительной вероятности P, объеме выборки n и значениях объясняющих переменных X указать такую предельную (гарантированную) величину погрешности $\delta_{P,n}(Y_{ср}(X))$, что

$$|Y_{ср}(X) - \widehat{\mathrm{f}}(X)| < \delta_{P,n}(Y_{ср}(X))$$

с вероятностью, не меньшей, чем P (здесь $Y_{ср}(X) = \mathbf{E}(\eta|X)$ — неизвестное *условное среднее* значение исследуемого результирующего показателя при значениях объясняющих переменных, равных X, а $\widehat{\mathrm{f}}(X)$ — построенная в соответствии с (В.24) эмпирическая функция регрессии);

3) при заданных доверительной вероятности P, объеме выборки n и значениях объясняющих переменных X указать такую предельную (гарантированную) величину погрешности $\delta_{P,n}(Y(X))$, что

$$Y(X) - \widehat{\mathrm{f}}(X)| \leqslant \delta_{P,n}(Y(X))$$

с вероятностью, не меньшей, чем P (здесь $Y(X)$ — прогнозируемое *индивидуальное* значение исследуемого результирующего показателя при значениях объясняющих переменных, равных X).

Описанию методов анализа точности исследуемых регрессионных моделей посвящена гл. 11 настоящего издания.

Заметим в заключение, что часть исследования, объединяющая этапы 4, 5, 6 и 7, принято называть *регрессионным анализом*.

ВЫВОДЫ

1. Аппарат статистического исследования зависимостей — составная часть многомерного статистического анализа — нацелен на решение основной проблемы естествознания: как на основании частных результатов статистического наблюдения за анализируемыми событиями или показателями выявить и описать существующие между ними стохастические взаимосвязи.

2. Анализируемые переменные величины по своей роли в исследовании подразделяются на результирующие (прогнозируемые) Y и объясняющие (предсказывающие, или предикторные) X. Среди компонент векторов Y и X могут быть и количественные, и порядковые (ординальные), и классификационные (номинальные).

3. Центральным математическим объектом в процессе статистического исследования зависимостей является функция $\mathfrak{f}(X)$, называемая функцией регрессии Y по X и описывающая, как правило[1], изменение условного среднего значения $Y_{ср}(X)$ результирующего показателя Y (вычисленного при фиксированных на уровне X значениях объясняющих переменных) в зависимости от изменения значений объясняющих переменных X.

4. Конечные прикладные цели статистического исследования зависимостей могут быть в основном трех типов: 1) установление самого факта наличия (или отсутствия) статистически значимой связи между Y и X, исследование структуры этих связей; 2) прогноз (восстановление) неизвестных значений индивидуальных или средних значений результирующего показателя по заданным значениям соответствующих объясняющих (предикторных) переменных; 3) выявление причинных связей между объясняющими переменными X и результирующими показателями Y, частичное управление значениями Y путем регулирования величин объясняющих переменных X.

5. Разделы многомерного статистического анализа, составляющие математический аппарат статистического исследования зависимостей, формировались и развивались с учетом специфики анализируемых моделей, обусловленной в первую очередь природой исследуемых переменных. Так, изучение зависимостей между количественными переменными обслуживается регрессионным и корреляционным анализами и анализом временных рядов (гл. 1—12, 14), изучение зависимостей количественного результирующего показателя от неколичественных или разнотипных объясняющих переменных — дисперсионным и ковариационным анализами, моделями типологической регрессии (гл. 13); для исследования зависимостей в условиях активного эксперимента служит теория оптимального планирования экспериментов [2, 3, 136]; наконец, для исследования системы зависимостей, в которых одни и те же

---

[1] В общей постановке задачи функция $\mathfrak{f}(X)$ может описывать поведение и других условных характеристик места группирования наблюдений результирующего признака $\eta(X)$, например условной медианы.

переменные в разных уравнениях этой системы могут одновременно выполнять и роль результирующих, и роль объясняющих, служит теория систем одновременных эконометрических уравнений (гл. 14). Аппарат исследования зависимостей неколичественных или разнотипных результирующих показателей от количественных или разнотипных объясняющих переменных в книге не рассматривается.

6. К основным типовым задачам практики, в которых использование аппарата статистического исследования зависимостей оказывается наиболее уместным и эффективным, следует отнести задачи: 1) нормирования; 2) прогноза, планирования и диагностики; 3) оценки труднодоступных (для непосредственного наблюдения и измерения) характеристик исследуемой системы; 4) оценки эффективности функционирования (или качества) анализируемой системы; 5) регулирования параметров функционирования анализируемой системы. Все эти задачи являются основными составными частями центральной проблемы кибернетики — проблемы «управления, связи и переработки информации» (см.: Математическая энциклопедия. Т. 2 — М.: Советская энциклопедия, 1979, с. 850).

7. По своей природе исследуемые зависимости могут быть разделены на: 1) детерминированные (тип A), когда исследуется функциональная зависимость между неслучайными переменными; 2) регрессионные (тип B), когда исследуется зависимость случайного результирующего показателя от неслучайных объясняющих переменных — параметров системы; 3) корреляционные (тип C), когда исследуется зависимость между случайными переменными, причем объясняющие переменные могут быть измерены без искажений; 4) конфлюэнтные (типы D_1 и D_2), когда исследуется функциональная зависимость между случайными или неслучайными переменными в ситуации, когда те и другие могут быть измерены только с некоторой случайной ошибкой.

8. Весь процесс статистического исследования зависимостей может быть разбит на семь последовательно реализуемых основных этапов, хронологический характер связей которых дополняется связями итерационного взаимодействия (см. рис. В.8): этап 1 (постановочный); этап 2 (информационный); этап 3 (корреляционный анализ); этап 4 (определение класса допустимых решений); этап 5 (анализ мультиколлинеарности предсказывающих переменных и отбор наиболее информативных из них); этап 6 (вычисление оценок неизвестных параметров, входящих в исследуемое уравнение статистической связи); этап 7 (анализ точности полученных уравнений связи).

Раздел I. АНАЛИЗ СТРУКТУРЫ И ТЕСНОТЫ СТАТИСТИЧЕСКОЙ СВЯЗИ МЕЖДУ ИССЛЕДУЕМЫМИ ПЕРЕМЕННЫМИ

(корреляционный анализ)

Имеется ли вообще какая-либо связь между исследуемыми переменными, какова структура этих связей и как измерить их тесноту? — эти вопросы исследователь ставит перед собой уже на ранней стадии статистического исследования зависимостей (см. описание этапа 3 в § В.6).

В частности, исследователь должен уметь: а) выбрать (с учетом специфики и природы анализируемых переменных) подходящий измеритель статистической связи (индекс или коэффициент корреляции, корреляционное отношение, какую-либо информационную характеристику связи, ранговый коэффициент корреляции и т. п.); б) оценить (с помощью точечной и интервальной оценок) его числовое значение по имеющимся выборочным данным; в) проверить гипотезу о том, что полученное числовое значение анализируемого измерителя связи действительно свидетельствует о наличии статистической связи (или, как говорят, проверить исследуемую корреляционную характеристику на статистически значимое ее отличие от нуля); г) проанализировать структуру связей между компонентами исследуемого многомерного признака, снабдив проведенный анализ специальным плоским геометрическим представлением исследуемой структуры, в котором компоненты (переменные) изображаются точками, а связи между ними — соединяющими их отрезками (см. рис. 4.1 и 4.2). Описанию методов и моделей, привлекаемых для решения всех тих вопросов, и посвящен данный раздел.

Глава 1. АНАЛИЗ ТЕСНОТЫ СВЯЗИ МЕЖДУ КОЛИЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

1.1. Анализ парных связей

1.1.1. Понятие индекса корреляции. Прежде чем приступать к исследованию конкретного вида связей между рассматриваемыми переменными, т. е. к оценке неизвестных параметров Θ в соотношениях типа

$$\mathbf{E}(\eta|X) = f(X;\ \Theta), \qquad (1.1)$$

следует выяснить, существует ли вообще эта связь, и, в случае положительного ответа, попытаться установить степень тесноты этой связи.

Во введении (§ В.5) описаны различные типы зависимостей, которые могут наблюдаться между исследуемыми переменными. Умение правильно классифицировать каждую конкретную многомерную систему наблюдений играет решающую роль при выборе соответствующих математико-статистических методов поиска изучаемой зависимости и при ее неформальной, физически содержательной интерпретации.

Однако в данном пункте в целях унификации подхода к решению исследуемой в этой главе задачи мы временно прибегнем к некоторому формальному обобщению рассмотренных ранее схем B, C и D. В частности, будет предложен подход, при котором во всех вышеупомянутых схемах зависимостей исследуемая независимая переменная интерпретируется как случайная переменная (параметр) ξ, от которой зависит закон условного распределения зависимой переменной η.

Итак, при каждом фиксированном значении $\xi = X$ распределение зависимой переменной $\eta(X)$ задается плотностью[1] $\varphi(Y|X)$, зависящей от X. Соответственно будут зависеть от X и математическое ожидание $\mathbf{E}\eta(X) = \mathbf{E}(\eta|X) = f(X)$, и дисперсия $\mathbf{D}\eta(X) = \mathbf{D}(\eta|X) = \sigma^2 h^2(X)$. Природа же исследуемой многомерной схемы, т. е. тип искомой зависимости, будет определяться спецификой частного закона распределения наблюдаемой независимой переменной ξ.

Очевидно, в схеме B (наблюдения производятся в *фиксированных* точках $X_1, \ldots, X_n$ без случайных ошибок в регистрации независимой переменной) случайную величину ξ следует рассматривать как дискретную с областью мыслимых значений $\mathbf{B} = \{X_1, X_2, \ldots, X_n\}$ (не исключается возможность повторения одинаковых значений ξ в этом ряду) и с частным законом распределения $\psi(X)$, задаваемым вероятностями $\psi(X_1) = \psi(X_2) = \ldots = \psi(X_n) = 1/n$.

В схеме D_1 плотность $\psi(X)$ частного распределения определяется, помимо набора наблюдаемых абсцисс $X_1, X_2, \ldots, X_n$, законами распределения ошибок измерения ε_X. Если k ($k \leqslant n$) — число различных уровней $X_1^0, X_2^0, \ldots, X_k^0$ структурной компоненты X, при которых снимались эксперимен-

---

[1]Все наши рассуждения остаются в силе и для дискретных случайных величин $\eta(X)$; при этом лишь надо заменить условные плотности $\varphi(Y|X)$ на соответствующие условные вероятности $\varphi(Y_i^0|X)$, где Y_i^0 — возможные значения исследуемого результирующего показателя.

тальные данные (X_1, y_1), (X_2, y_2), ..., (X_n, y_n), а $\psi_i(X)$ — плотность распределения ошибки $\varepsilon_{X_i^0}$, то

$$\psi(X) = \sum_{i=1}^{k} \frac{n_i}{n} \psi_i(X - X_i^0),$$

где n_i — число наблюдений, произведенных «на уровне» $X = X_i^0$.

В схемах C и D_2 объясняющие наблюдаемые переменные соответственно ξ и ξ' по своей природе случайны, следовательно, им также соответствует некоторая плотность частного распределения $\psi(X)$.

Если рассмотреть случай *единственного* результирующего показателя η и мысленно спроектировать все точки исследуемой многомерной системы на ось его возможных значений Oy, то получим выборку из одномерного закона с плотностью $\varphi(y)$, характеризующего вероятностную природу безусловной случайной величины η. При такой интерпретации очевидно, что плотность частного (безусловного) распределения $\varphi(y)$ получается как смесь соответствующих условных плотностей $\varphi(y|X)$, а именно: $\varphi(y) = \int_{\mathbf{B}} \varphi(y|X)\psi(X)\,dX$ (в схеме B $\varphi(y) = \sum_{i=1}^{n} \varphi(y|X_i)\psi(X_i)$; в дальнейшем при усреднении по $\psi(X)$ мы не будем специально оговаривать случай схемы B, подразумевая переход от интегрирования по X к суммированию по X_i). Соответственно в нашем дальнейшем изложении будут участвовать характеристики

$$m_\eta = \mathbf{E}\eta = \int_{\mathbf{B}} f(X)\psi(X)\,dX;$$

$$\sigma_\eta^2 = \mathbf{D}\eta = \int_{\mathbf{Y}} (y - m_\eta)^2 \varphi(y)\,dy;$$

$$\sigma_f^2 = \mathbf{D}f(\xi) \doteq \int_{\mathbf{B}} (f(X) - \mathbf{E}f(\xi))^2 \psi(X)\,dX;$$

$$\bar{\sigma}_{\eta(X)}^2 = \mathbf{E}[\sigma^2 h^2(\xi)] = \int_{\mathbf{B}} \sigma^2 h^2(X)\psi(X)\,dX.$$

Рассмотрим, например, частный случай схемы C, когда вектор исследуемых показателей

$$\begin{pmatrix}\xi\\ \eta\end{pmatrix} = (\xi^{(1)}, \ldots, \xi^{(p)}; \eta^{(1)}, \ldots, \eta^{(m)})' \qquad (1.2)$$

— $(p+m)$-мерная нормальная случайная величина [14, с. 173], и пусть $\mathbf{M}_\xi = \mathbf{E}\xi$, $\mathbf{M}_\eta = \mathbf{E}\eta$ — соответственно векторы средних значений объясняющих переменных $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})'$ и результирующих показателей $\eta = (\eta^{(1)}, \ldots, \eta^{(m)})'$, а $\Sigma_{\xi\xi}$, $\Sigma_{\xi\eta}$ и $\Sigma_{\eta\eta}$ — ковариационные матрицы [14, с. 138] соответственно векторов ξ, ξ и η, η[1]. Тогда можно показать (см., например, [20, с. 45]), что условное распределение вектора результирующих показателей $\eta = (\eta^{(1)}, \ldots, \eta^{(m)})'$ при условии, что значения объясняющих переменных зафиксированы на уровне $X = (x^{(1)}, \ldots, x^{(p)})'$ (т. е. при условии $\xi = X$), также нормально с условным средним значением

$$\mathbf{E}(\eta \mid \xi = X) = \mathbf{M}_\eta + \Sigma_{\eta\xi}\Sigma_{\xi\xi}^{-1}(X - \mathbf{M}_\xi) \tag{1.3}$$

и ковариационной матрицей

$$\mathbf{E}\{(\eta - \mathbf{M}_\eta)(\eta - \mathbf{M}_\eta)' \mid \xi = X\} = \Sigma_{\eta\eta} - \Sigma_{\eta\xi}\Sigma_{\xi\xi}^{-1}\Sigma_{\xi\eta}. \tag{1.4}$$

Из (1.3) и (1.4), в частности, следует:

а) функция $\mathbf{f}(X) = \mathbf{E}(\eta \mid \xi = X)$ регрессии η по ξ при совместном нормальном законе распределения исследуемых показателей линейна по X;

б) ковариационная матрица условного распределения вектора результирующих показателей $\eta(X) = (\eta \mid \xi = X)$ не зависит от X;

в) если рассматривается *парная* регрессионная зависимость, т. е. зависимость единственного результирующего показателя η от единственной объясняющей переменной ξ в схеме C, причем распределение случайной величины (ξ, η) подчиняется двумерному нормальному закону, то условное распределение случайной величины $\eta(x) = (\eta \mid \xi = x)$ тоже нормально с условным средним значением (функцией регрессии)

$$\mathbf{E}(\eta \mid \xi = x) = f(x) = m_\eta + r\frac{\sigma_\eta}{\sigma_\xi}(x - m_\xi) \tag{1.3'}$$

и с дисперсией

$$\mathbf{D}(\eta \mid \xi = x) = \sigma_\eta^2(1 - r^2) \tag{1.4'}$$

[1] Ковариационные матрицы $\Sigma_{\xi\xi}$, $\Sigma_{\xi\eta}$, $\Sigma_{\eta\xi}$ и $\Sigma_{\eta\eta}$ получаются из общей ковариационной матрицы Σ вектора $\binom{\xi}{\eta}$ ее разбиением на блоки по следующей схеме: $\Sigma = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\eta} \\ \Sigma_{\eta\xi} & \Sigma_{\eta\eta} \end{pmatrix}$. Соответственно размерности матриц $\Sigma_{\xi\xi}$, $\Sigma_{\xi\eta}$, $\Sigma_{\eta\xi}$ и $\Sigma_{\eta\eta}$ будут: $p \times p$, $p \times m$, $m \times p$ и $m \times m$.

(здесь m_ξ и m_η — средние значения соответственно объясняющей переменной ξ и результирующего показателя η, σ_ξ^2 и σ_η^2 — их дисперсии, а r — коэффициент корреляции между ними, см., например, [14, гл. 5]).

Будем рассматривать в дальнейшем (если специально не оговорено противное) случай *единственного* результирующего показателя, т. е. случай $m = 1$.

Итак, величина $\sigma_\eta^2 = \mathbf{D}\eta$ характеризует полную вариацию (дисперсию) исследуемого результирующего показателя η, в то время как $\sigma_f^2 = \mathbf{D}f(\xi)$ определяет дисперсию функции регрессии $y_{\text{ср}} = f(x)$, $\overline{\sigma}_{\eta(x)}^2$ — усредненную (по различным значениям ξ) величину условной дисперсии $\mathbf{D}(\eta \mid \xi = x)$, т. е. среднюю величину дисперсии неконтролируемой остаточной случайной компоненты ε (см. соотношения (В.14), (В.16), (В.21)).

Воспользовавшись соотношением (2в.3.6) из [117, с. 94], получим следующее полезное соотношение, связывающее три вышеупомянутые меры случайного разброса:

$$\sigma_\eta^2 = \sigma_f^2 + \overline{\sigma}_{\eta(x)}^2. \tag{1.5}$$

Это означает, что полная вариация исследуемой зависимой переменной складывается из контролируемой нами вариации функции регрессии и из не поддающейся нашему контролю вариации остаточной случайной компоненты. Очевидно, связь между η и ξ в соотношениях (В.14), (В.16), (В.21) и т. п. будет тем теснее, тем определеннее, чем менее «размазанными» окажутся участвующие в них остаточные неконтролируемые случайные компоненты $\varepsilon(X)$, ε и $\varepsilon_y + \sum_{\nu=1}^{\infty} [(-\varepsilon_x)^\nu/\nu!]\, f^{[\nu]}(\xi)$. Можно, в частности, задаться вопросом: какая доля степени изменчивости интересующего нас зависимого признака (т. е. какая доля дисперсии σ_η^2) обусловливается изменчивостью описывающей его функции независимой переменной $f(\xi)$ (т. е. ее дисперсией σ_f^2)? Так мы приходим к понятию наиболее общей характеристики степени тесноты связи между η и ξ — *индекса корреляции* $I_{\eta \cdot \xi}$, где

$$I_{\eta \cdot \xi}^2 = \frac{\sigma_f^2}{\sigma_\eta^2} = 1 - \frac{\overline{\sigma}_{\eta(x)}^2}{\sigma_\eta^2}. \tag{1.6}$$

Из (1.5) и (1.6) непосредственно следует, что $0 \leqslant I_{\eta \cdot \xi} \leqslant 1$. При этом минимальное значение индекса корреляции ($I_{\eta \cdot \xi} = 0$) соответствует полному отсутствию варьирования $f(\xi)$

с изменением ξ ($\sigma_f = 0$), а это означает полное отсутствие какого-либо влияния ξ на η, т. е., как говорят, отсутствие корреляционной связи между результирующим показателем η и объясняющими переменными ξ.

В то же время максимальное значение индекса корреляции ($I_{\eta\cdot\xi} = 1$) соответствует полному отсутствию варьирования остаточной случайной компоненты ($\sigma_{\eta(x)} = 0$). А поскольку среднее значение остаточной случайной компоненты равно нулю, то она практически исчезает из разложений (В.14), (В.16), (В.21). Это означает наличие чисто функциональной связи между η и ξ и, следовательно, возможность детерминированного восстановления значений η по соответствующим значениям объясняющих переменных ξ.

Таким образом, введенный с помощью (1.6) индекс корреляции $I_{\eta\cdot\xi}$ между результи-рующим показателем η и объясняющими переменными ξ формально определен для любой двумерной системы наблюдений. Квадрат его величины ($I^2_{\eta\cdot\xi}$) показывает, какая доля дисперсии исследуемого результирующего показателя η определяется (детерминируется) изменчивостью (дисперсией) соответствующей функции регрессии f от аргумента ξ, поэтому часто называется коэффициентом детерминации. Соответственно оставшаяся доля дисперсии η (т. е. $1 - I^2_{\eta\cdot\xi}$) объясняется воздействием неконтролируемой случайной остаточной компоненты («помехи»), а следовательно, определяет ту верхнюю границу точности, с которой мы сможем восстанавливать (предсказывать) значения η по заданным значениям объясняющих переменных ξ.

Наилучшие методы построения статистической оценки $\widehat{I}_{\eta\cdot\xi}$ для неизвестного теоретического значения индекса корреляции $I_{\eta\cdot\xi}$, так же как и различные варианты его интерпретации, зависят от ряда исходных предпосылок каждой конкретной двумерной схемы (общий вид функции $f(\xi)$, вид распределения многомерной случайной величины (ξ, η) и т. п.). Описание их поэтому дается ниже отдельно для каждого из некоторых специальных частных случаев.

1.1.2. Коэффициент корреляции как измеритель степени тесноты связи в двумерных нормальных схемах. Пусть исследуется парная зависимость между случайными переменными η и ξ типа C (или между η' и ξ' типа D), см. § В.5. Предположим, что имеющиеся в нашем распоряжении результаты наблюдения (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) представляют собой выборку из двумерной нормальной генеральной совокупности (см. [14, с. 171]) В этом случае введенный ранее (1.6) индекс корреляции просто выражается через коэффициент корреля-

ции r, участвующий в записи уравнения соответствующей двумерной нормальной плотности. Воспользовавшись соотношением (1.6) с учетом (1.4′), получаем

$$I_{\eta \cdot \xi} = r. \tag{1.7}$$

С помощью непосредственных вычислений, опирающихся на формулу для плотности двумерного нормального закона, можно показать, что

$$r = \frac{\mathbf{E}\,[(\xi - \mathbf{E}\,\xi)(\eta - \mathbf{E}\eta)]}{\sqrt{\mathbf{D}\xi \cdot \mathbf{D}\eta}} = \frac{\operatorname{cov}(\xi,\ \eta)}{\sigma_\xi \cdot \sigma_\eta}, \tag{1.8}$$

где ковариация cov (ξ, η) — второй центральный смешанный момент двумерной случайной величины (ξ, η), а σ_ξ и σ_η — среднеквадратические (безусловные) отклонения соответственно компонент ξ и η. Величина r, определенная соотношением (1.8), называется *коэффициентом корреляции*[1] и характеризует (в силу (1.7)) степень тесноты связи между случайными компонентами ξ и η. При этом лишь в данном частном случае характеристика степени тесноты связи симметрична относительно переменных ξ и η (т. е. $r_{\xi\eta} = r_{\eta\xi}$) и имеет поддающийся содержательной интерпретации знак «плюс» или «минус». Положительность коэффициента корреляции r означает одинаковый характер тенденции взаимосвязанного изменения случайных компонент ξ и η: с увеличением ξ мы наблюдаем тенденцию увеличения соответствующих индивидуальных значений η и, следовательно, увеличивается условное математическое ожидание $\mathbf{E}\,(\eta \mid \xi = x)$. Отрицательное значение r говорит о противоположной тенденции взаимосвязанного изменения компонент ξ и η (с увеличением ξ уменьшается $\mathbf{E}\,(\eta \mid \xi = x)$).

Выборочное значение $\hat{r}$ *коэффициента корреляции* (т. е. статистическая оценка $\hat{r}$ неизвестного значения r) подсчитывается по исходным статистическим данным $(x_1,\ y_1)$, $(x_2,\ y_2)$, …, $(x_n,\ y_n)$ по формуле

$$\hat{r} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{1.8′}$$

[1] В ситуациях, когда наряду с r рассматриваются *частные* (см. п. 1.2.2) и *множественные* (см. п. 1.3.2) коэффициенты корреляции, его называют *парным коэффициентом корреляции.*

где $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ и $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

Определенные соотношениями (1.8) и (1.8′) соответственно теоретический и выборочный коэффициенты корреляции могут быть формально вычислены для любой двумерной системы наблюдений; они являются измерителями степени тесноты *линейной* статистической связи между анализируемыми признаками. Однако только в случае совместной нормальной распределенности исследуемых случайных величин ξ и η коэффициент корреляции r имеет четкий смысл как характеристика степени тесноты связи между ними. В частности, в этом случае соотношение $|r| = 1$ подтверждает чисто функциональную линейную зависимость между исследуемыми величинами, а уравнение $r = 0$ свидетельствует об их полной взаимной независимости. Кроме того, коэффициент корреляции вместе со средними и дисперсиями случайных величин ξ и η составляет те пять параметров, которые дают исчерпывающие сведения о стохастической зависимости исследуемых величин, так как однозначно определяют их двумерный закон распределения (см. [14, с. 171, формула (6.9)]).

Во всех же остальных случаях (распределения ξ и η отклоняются от нормального, одна из исследуемых величин не является случайной и т. п.) коэффициент корреляции можно использовать лишь в качестве одной из возможных характеристик степени тесноты связи. При этом, несмотря на то, что в общем случае пока не предложено характеристики линейной связи, которая обладала бы очевидными преимуществами по сравнению с r, его интерпретация часто оказывается весьма ненадежной. Если же априори допускается возможность отклонения от линейного вида зависимости, то можно построить примеры, когда, несмотря на $r = 0$, исследуемые переменные оказываются связанными чисто функциональным соотношением (следовательно, $I^2 = 1$). Поэтому о величинах, для которых $r = 0$, обычно говорят, что они *некоррелированы*, и только после дополнительного статистического и профессионального анализа (исследование степени отклонения распределения рассматриваемых величин от нормального и т. п.) можно сказать, следует ли отсюда их *независимость*. И, наоборот, из высокой степени коррелированности величин при сильных отклонениях распределения ξ и η от нормального еще не следует их столь же тесная зависимость.

Приведем пример. На рис. 1.1, *а* представлены данные, характеризующие численность населения ξ и соответствую-

щее число телевизионных точек η в девяти городах США — Денвере, Сан-Антонио, Канзас-Сити, Сиэтле, Цинциннати, Буффало, Нью-Орлеане, Милуоки, Хьюстоне[1].

По формуле (1.8′) получаем, что коэффициент корреляции $\hat{r} = 0{,}403$; это при $n = 9$ свидетельствует о весьма малой степени коррелированности ξ и η. Если же к этим данным присовокупить соответствующие сведения о Нью-Йорке ($x_{10} = 802$; $y_{10} = 345$, см. рис. 1.1, *б*), то объем выборки увеличивается

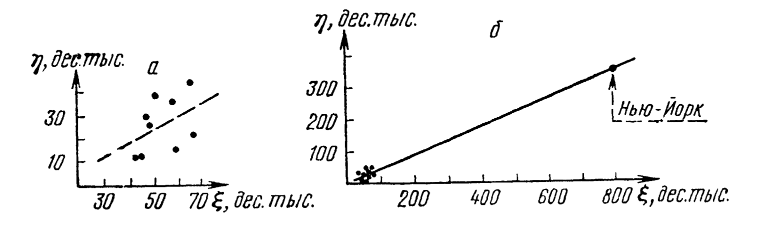


Рис. 1.1. Корреляционное поле, характеризующее связь между численностью населения ξ и числом установленных телевизионных точек η в США в 1953 г.:
а) в девяти городах; б) в десяти городах

на единицу ($n' = 10$), а соответственно пересчитанный коэффициент корреляции $\hat{r} = 0{,}995$. Дело здесь в том, что последнее (десятое) наблюдение является «аномальным», резко выделяющимся, так что всю совокупность наблюдений мы уже не можем считать выборкой из одной и той же нормальной генеральной совокупности (в чем читатель сможет без труда убедиться, воспользовавшись одним из приемов, описанных в [14, § 11.5]).

И наконец, даже если удалось установить тесную зависимость между двумя исследуемыми величинами, отсюда еще непосредственно не следует их *причинная взаимообусловленность*. Например, при анализе большого числа наблюдений, относящихся к отливке труб на сталелитейных заводах, была установлена положительная корреляционная связь между временем плавки и процентом забракованных труб [10]. Дать какое-либо причинное истолкование этой стохастической связи было невозможно, а поэтому рекомендации ограничить продолжительность плавки для снижения процента забракованных труб выглядели малосостоятельными. Действительно, спустя несколько лет обнаружили, что большая продолжительность

---

[1]См.: Миллс Ф. Статистические методы. — М.: Госстатиздат, 1958. — 799 с.

плавки всегда была связана с использованием сырья специального состава. Этот вид сырья приводил одновременно к длительному времени плавки и большому проценту брака, хотя оба эти фактора взаимно независимы.

Таким образом, высокий коэффициент корреляции между продолжительностью плавки и процентом забракованных труб полностью обусловливался влиянием третьего, не учтенного при исследовании фактора — характеристики качества сырья. Если же этот фактор был бы с самого начала учтен, то никакой значимой корреляционной связи между временем плавки и процентом забракованных труб мы бы не обнаружили. За счет подобных эффектов (одновременного влияния неучтенных факторов на исследуемые переменные) может ис кажаться и смысл истинной связи между переменными, т. е., например, подсчеты приводят к положительному значению парного коэффициента корреляции, в то время как истинная связь между ними имеет отрицательный смысл. Такую корреляцию между двумя переменными часто называют «ложной». Более детально подобные ситуации — обнаружение и исключение «общих причинных факторов», расчет «очищенных», или *частных*, коэффициентов корреляции и т. п. — исследуют методами многомерного корреляционного анализа (см. § 1.2). Такого рода недоразумения с причинным толкованием статистических связей наиболее вероятны в ситуациях, когда исходными статистическими данными являются показатели работы действующего предприятия. Их обычно удается свести к минимуму при получении данных из искусственно поставленного эксперимента.

Выборочное значение коэффициента корреляции в примере В.3 между процентом забракованного стекла и соответствующей величиной вакуума в печи для его отжига $\widehat{r} = -0{,}655$[1]. Оно, по-видимому, свидетельствует о наличии определенной зависимости между исследуемыми переменными. Однако утверждать, что повышение вакуума в печи причинно обусловливает понижение процента брака, преждевременно: предварительно следует провести дополнительный профессионально-

---

[1]При подсчетах, связанных с примером В.3, здесь и в дальнейшем пользуемся данными, соответствующими рис. В.6. При этом диапазон независимой переменной (т . е. диапазон значений вакуума в печи) разбиваем на шесть равных интервалов шириной 0,5: $n = 43$; $k = 6$; $m_1 = 3$; $m_2 = 4$; $m_3 = 8$; $m_4 = 13$; $m_5 = 11$; $m_6 = 4$. Далее воспользуемся и разбиением диапазона зависимой переменной η (процента забракованного стекла): от 1,5 через 0,5 до 4,5%, так что и число интервалов группирования по вертикальной оси (k_y) в данном случае также равно шести.

статистический анализ, в частности выяснить, нет ли в технологических условиях данного эксперимента неучтенного фактора, изменения которого одновременно приводили бы к повышению вакуума и понижению брака производства.

Замечания о необходимости известной осторожности при толковании корреляционной связи никоим образом не обесценивают желательность проверки значимости любого кажущегося соотношения. При этом следует использовать характеристики степени тесноты связи: коэффициента корреляции $\hat{r}$ и корреляционного отношения ρ (см. ниже). Но не всегда знание этих характеристик оказывается достаточным для получения информации о степени тесноты физической связи между исследуемыми переменными и тем более об их причинной взаимообусловленности.

1.1.3. Распределение выборочного коэффициента корреляции и проверка гипотезы о статистической значимости линейной связи. Какую величину выборочного коэффициента корреляции следует считать достаточной для статистически обоснованного вывода о наличии корреляционной связи между исследуемыми переменными? Ведь надежность статистических характеристик, в том числе и $\hat{r}$, ослабевает с уменьшением объема соответствующей выборки, а потому принципиально возможны случаи, когда отклонение от нуля полученной величины выборочного коэффициента корреляции $\hat{r}$ оказывается статистически незначимым, т. е. целиком обусловленным неизбежным случайным колебанием выборки, на основании которой он вычислен. Ответить на этот вопрос помогает знание закона вероятностного распределения $\hat{r}$. В случае совместной нормальной распределенности исследуемых переменных и при достаточно большом объеме выборки n распределение $\hat{r}$ можно считать приближенно нормальным со средним, равным своему теоретическому значению r и дисперсией $\sigma_{\hat{r}}^2 = \frac{(1-r^2)^2}{n}$ [10, с. 104]. Однако следует учитывать, что при малых значениях n и r, близких к ± 1, это приближение оказывается очень грубым. Кроме того, при малых n следует принимать во внимание, что величина $\hat{r}$ является *смещенной* оценкой своего теоретического значения r, в частности $\mathbf{E}\hat{r} = r - [r(1 - r^2)]/2n$.

Относительно хорошая степень приближения нормального распределения при малых значениях $|r|$ позволяет получить простой критерий проверки гипотезы $r = 0$, т. е. гипотезы об отсутствии корреляционной связи между исследуе-

мыми переменными. Используется тот факт, что величина $t^{(\hat r)}(n-2)=\dfrac{\hat r\sqrt{n-2}}{\sqrt{1-\hat r^2}}$ при условии $r=0$ распределена по закону Стьюдента с $n-2$ степенями свободы (см., например, [117, с. 181]. Поэтому если окажется, что

$$\frac{|\hat r|\sqrt{n-2}}{\sqrt{1-\hat r^2}}<t_{0,05}(n-2) \qquad (1.9)$$

(здесь $t_{0,05}(n-2)$ — 5%-ная точка распределения Стьюдента с $n-2$ степенями свободы), то гипотеза об отсутствии корреляционной связи принимается. Используем этот критерий для исследования значимости корреляционной связи в примере В.3: $\widehat{t^{(r)}}=7{,}3$; $t_{0,05}(41)=1{,}68$; так что гипотеза об отсутствии корреляционной связи между процентом забракованного стекла и вакуумом в печи для его отжига должна быть отвергнута.

Доверительные интервалы для истинного значения коэффициента корреляции r можно построить из нормальной распределенности $\hat r$. Концы интервала $[r_1, r_2]$ можно вычислять по приближенной формуле

$$r_{1.2}\approx\hat r+\frac{\hat r(1-\hat r^2)}{2n}\mp u_{\frac{\alpha}{2}}\frac{1-\hat r^2}{\sqrt n}. \qquad (1.10)$$

Здесь $u_{\alpha/2}$, в соответствии с ранее введенными обозначениями, — $100\frac{\alpha}{2}$ %-ная точка стандартного (0, 1)-нормального распределения, так что истинное значение коэффициента корреляции r с доверительной вероятностью $1-\alpha$ принадлежит интервалу $[r_1, r_2]$. Однако использование формулы (1.10) сопряжено со следующими оговорками: истинное значение коэффициента корреляции не должно быть близким к ± 1; общее число наблюдений n должно быть достаточно велико; величина r в поправке к «смещению» r (т. е. в выражении $\frac{r(1-r^2)}{2n}$) и в дисперсии $\sigma^2_{\hat r}$ заменена ее приближенным (выборочным) значением $\hat r$. Избавиться от этих ограничений позволяет следующее преобразование, предложенное Р. Фишером (см., например, [117, с. 383]:

$$z=\frac{1}{2}\ln\frac{1+\hat r}{1-\hat r}. \qquad (1.11)$$

Он показал, что величина z, определенная соотношением (1.11), уже при небольших n с хорошим приближением следует нормальному закону со средним $\mathbf{E}z \approx \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{r}{2(n-1)}$ и дисперсией $\mathbf{D}z = \frac{1}{n-3}$. Это позволяет построить доверительный интервал $[z_1, z_2]$ для $\mathbf{E}z$ по формуле

$$z_{1,2} = \frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}} \mp \frac{u_{\frac{\alpha}{2}}}{\sqrt{n-3}} - \frac{\hat{r}}{2(n-1)} =$$

$$= \operatorname{arcth}\hat{r} \mp \frac{u_{\frac{\alpha}{2}}}{\sqrt{n-3}} - \frac{\hat{r}}{2(n-1)},$$

откуда следует, что истинное значение коэффициента корреляции r с той же доверительной вероятностью $1-\alpha$ заключено в пределах

$$\operatorname{th} z_1 < r < \operatorname{th} z_2. \qquad (1.12)$$

Здесь $\operatorname{th} z$ — это тангенс гиперболический от аргумента z (определяется с помощью соотношения $\operatorname{th} z = (e^z - e^{-z})/(e^z + e^{-z})$). Соответственно функция, определяющая величину z с помощью соотношения (1.11), это есть функция, обратная к тангенсу гиперболическому; так что часто вместо $z = \frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}}$ пишут $z = \operatorname{arcth}\hat{r}$ (или $z = \operatorname{th}^{-1}\hat{r}$). Нахождение z по данному значению $\hat{r}$ и, наоборот, определение $\hat{r}$ по заданной величине z производится с помощью табл. П. 7, в которой в крайних столбцах (левом и правом) приведены значения $|r|$, а между ними — соответствующие значения $|z| = \operatorname{arcth}|\hat{r}|$ (знаки у аргумента и функции совпадают, так что если, например, $\hat{r}$ отрицателен, то и соответствующее значение $z = \operatorname{arcth}\hat{r}$ также отрицательно).

Так, задавшись 95%-ной доверительной вероятностью в примере В.3, находим

$$z_1 = \operatorname{arcth}(-0{,}655) - \frac{1{,}96}{40} - \frac{-0{,}655}{2{,}42} = -1{,}07;$$

$$z_2 = \operatorname{arcth}(-0{,}655) + \frac{1{,}96}{40} - \frac{-0{,}655}{2{,}42} = -0{,}47.$$

С помощью табл. П.7 находим: $\operatorname{th} z_1 = -0{,}79$; $\operatorname{th} z_2 = -0{,}44$, так что с вероятностью 0,95 имеем $-0{,}79 < r < -0{,}44$.

Использование z — преобразованной величины r — оказывается более предпочтительным и при проверке значимости корреляционной связи, когда число наблюдений n мало.

При построении доверительных интервалов для коэффициента корреляции, так же как и при проверке статистической значимости корреляционной связи, можно пользоваться специальными таблицами и графиками, в частности номограммой, изображенной на рис. 1.2.

Так, для построения доверительного интервала с помощью приведенных на рис. 1.2 номограмм следует отложить значение выборочного коэффициента корреляции $\widehat{r}$ на горизонтальной оси и провести через эту точку вертикальную прямую. Ординаты r_1 и r_2 ($r_1 < r_2$) пересечения этой вертикальной прямой с двумя кривыми, над которыми надписан объем используемой выборки, и являются граничными точками искомого доверительного интервала, т. е. $r_1 < r < r_2$. Рис. 1.2, *а* дает решение поставленной задачи с доверительной вероятностью $P = 0{,}95$, а рис. 1.2, *б* — с доверительной вероятностью $P = 0{,}99$

Критерий однородности двух или нескольких выборочных коэффициентов корреляции. Пусть по выборкам объемов n_1 и n_2 из каких-то двух нормальных генеральных совокупностей получены выборочные значения коэффициентов корреляции $\widehat{r^{(1)}}$ и $\widehat{r^{(2)}}$. Можно ли признать различие в значениях $\widehat{r^{(1)}}$ и $\widehat{r^{(2)}}$ статистически значимым или же это различие обусловлено случайными колебаниями выборок, следовательно, полученные величины $\widehat{r^{(1)}}$ и $\widehat{r^{(2)}}$ не противоречат гипотезе о том, что две рассмотренные генеральные совокупности имеют один и тот же теоретический коэффициент корреляции?

Для статистической проверки этих предположений используется факт приближенной (0, 1)-нормальной распределенности статистики (справедливый лишь в предположении истинности гипотезы об однородности $\widehat{r^{(1)}}$ и $\widehat{r^{(2)}}$)

$$\gamma = (z^{(1)} - z^{(2)}) \Big/ \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

где $z^{(i)}$ ($i = 1, 2$) подсчитываются по формуле (1.11) соответственно отдельно по первой ($i = 1$) и по второй ($i = 2$) выборкам.

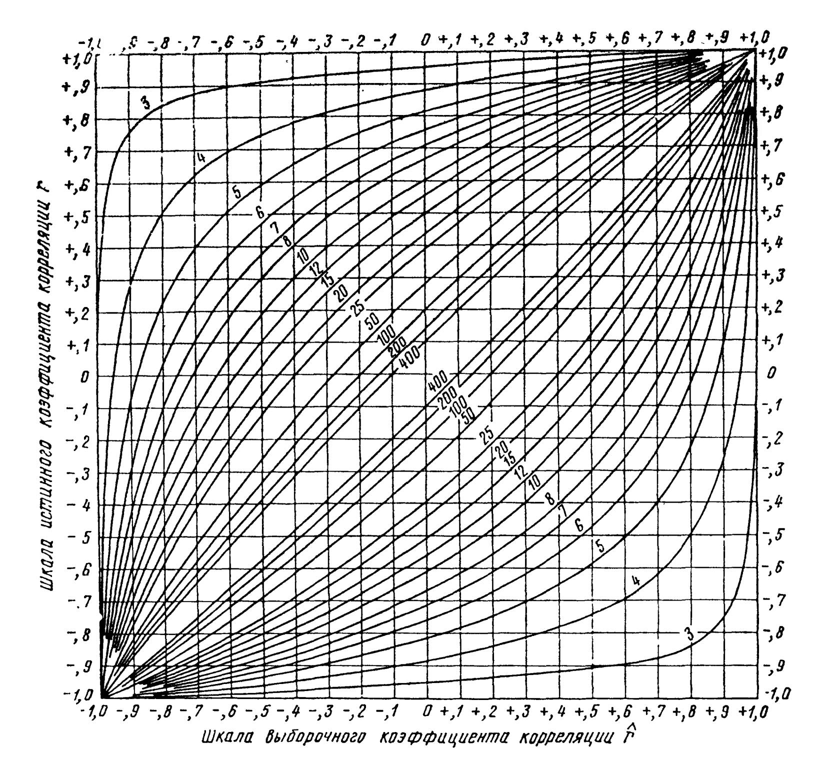
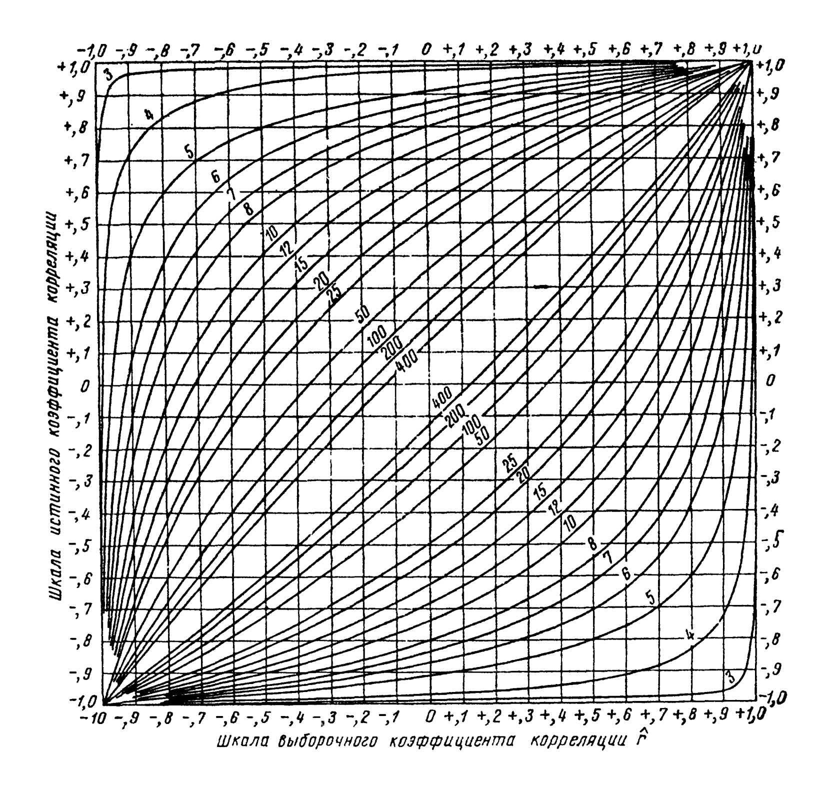


Рис. 1.2. Номограмма для построения доверительных интервалов для выборок с доверительной вероятностью: а) $P=0{,}95$; б) $P=0{,}99$

В частности, если оказалось, что $|\gamma| > u_{\frac{\alpha}{2}}$, то различие между $\widehat{r}^{(1)}$ и $\widehat{r}^{(2)}$ признается статистически значимым (с уровнем значимости α).

Пусть теперь $\widehat{r}^{(1)}, \widehat{r}^{(2)}, \ldots, \widehat{r}^{(k)}$ — k коэффициентов корреляции, полученных по выборкам объемов $n_1, n_2, \ldots, n_k$ из k каких-то нормальных генеральных совокупностей. Можно ли считать, что, несмотря на видимые различия в значениях выборочных коэффициентов корреляции $\widehat{r}^{(1)}, \widehat{r}^{(2)}, \ldots, \widehat{r}^{(k)}$, значение теоретического коэффициента корреляции r остается одним и тем же во всех обследованных генеральных совокупностях? Если допустить справедливость утвердительного ответа на поставленный вопрос, то статистика



истинного значения коэффициента корреляции при различных объемах

$$\widehat{\chi^2} = \sum_{i=1}^{k} (n_i - 3)\, z^{(i)^2} - \frac{\left(\sum\limits_{i=1}^{k} (n_i - 3)\, z^{(i)}\right)^2}{\sum\limits_{i=1}^{k} (n_i - 3)} \qquad (1.13)$$

должна приближенно подчиняться χ^2-распределению с $k - 1$ степенью свободы (здесь $z^{(i)}$, как и прежде, подсчитываются отдельно по каждой отдельной выборке по формуле (1.11)). Поэтому если окажется, что подсчитанное по формуле (1.13) значение $\widehat{\chi^2} > \chi^2_\alpha\,(k - 1)$, где $\chi^2_\alpha\,(k - 1)$ — величина 100α %-ной точки χ^2-распределения с $k - 1$ степенью свободы (см. табл. П.4), то гипотеза об однородности выборочных коэффи-

циентов корреляции $\widehat{r}^{(1)}, \widehat{r}^{(2)}, \ldots, \widehat{r}^{(k)}$ отвергается (с уровнем значимости α).

В табл. 1.1 приводится пример вычислений по схеме описанной процедуры (заимствован из [117, с. 386]).

Значение $\widehat{\chi}^2 = 4{,}4995 - \frac{(18{,}231)^2}{95} = 1{,}0009$ в данном примере оказалось существенно меньше 5%-ной точки χ^2-распределения с пятью степенями свободы ($\chi^2_{0,05}(5) = 11{,}07$), так что следует признать непротиворечивость полученных выборочных значений коэффициентов корреляции (0,318; 0,106; 0,253; 0,340; 0,116 и 0,112) с гипотезой об их однородности.

Таблица 1.1

Номер выборки (i)	Объем выборки минус три $(n_i - 3)$	Выборочный коэффициент корреляции $\widehat{r}^{(i)}$	$z^{(i)}$	$(n-3)\,z^{(i)}$	$(n-3)\,z^{(i)2}$
1	7	0,318	0,3294	2,3058	0,7595
2	11	0,106	0,1064	1,1704	0,1245
3	13	0,253	0,2586	3,3618	0,8694
4	17	0,340	0,3541	6,0197	2,1316
5	22	0,116	0,1164	2,5608	0,2981
6	25	0,112	0,1125	2,8125	0,3164
Сумма	95			18,2310	4,4995

1.1.4. Влияние ошибок измерения на величину коэффициента корреляции. Пусть мы хотим оценить степень тесноты корреляционной связи между компонентами двумерной нормальной случайной величины (ξ, η), однако наблюдать мы их можем лишь с некоторыми случайными «ошибками измерения» соответственно ε_ξ и ε_η (см. схему зависимости D_2 во введении). Поэтому экспериментальные данные (x_i, y_i), $i = 1, 2, \ldots, n$, — это практически выборочные значения искаженной двумерной случайной величины (ξ', η'), где $\xi' = \xi + \varepsilon_\xi$ и $\eta = \eta + \varepsilon_\eta$. Если предположить, что ε_ξ и ε_η взаимно независимы, не зависят от ξ и η, нормальны, имеют нулевые математические ожидания и конечные дисперсии соответственно σ_1^2 и σ_2^2, то двумерная случайная величина (ξ', η') будет также подчиняться двумерному нормальному распределению. Однако, как легко подсчитать, параметры этого распределения и, в частности коэффициент корреляции r' между ξ'

и η' будут соответственно отличаться от параметров исходной двумерной схемы (ξ, η). Действительно, в соответствии с основными правилами вычисления первых и вторых моментов [14, гл. 5] получаем:

$$\begin{aligned} a_{\xi'} &= \mathbf{E}\xi' = \mathbf{E}\xi = a_\xi; \\ a_{\eta'} &= \mathbf{E}\eta' = \mathbf{E}\eta = a_\eta; \\ \sigma^2_{\xi'} &= \sigma^2_\xi + \sigma^2_1; \\ \sigma^2_{\eta'} &= \sigma^2_\eta + \sigma^2_2; \\ r' &= \frac{r}{\sqrt{\left(1+\frac{\sigma_1^2}{\sigma_\xi^2}\right)\left(1+\frac{\sigma_2^2}{\sigma_\eta^2}\right)}}. \end{aligned} \tag{1.14}$$

Из (1.14), в частности, следует, что коэффициент корреляции признаков, на которые наложены ошибки измерения, всегда меньше по абсолютной величине, чем коэффициент корреляции исходных признаков. Другими словами, ошибки измерения всегда ослабляют исследуемую корреляционную связь между исходными переменными, и это искажение тем меньше, чем меньше отношения дисперсий ошибок к дисперсиям самих исходных переменных. Формула (1.14) позволяет скорректировать искаженное значение коэффициента корреляции: для этого нужно либо знать «разрешающие» характеристики измерительных приборов (и, следовательно, величины дисперсий ошибок σ_1^2 и σ_2^2), либо провести дополнительное исследование по их выявлению.

1.1.5. Измерение степени тесноты связи при нелинейной зависимости. При отклонениях исследуемой зависимости от линейного вида, как уже отмечалось, коэффициент корреляции r теряет свой смысл как характеристика степени тесноты связи. В этих случаях исследователь должен воспользоваться имеющимися у него двумерными выборочными данными (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) с целью построения оценок для определенной выше, в некотором смысле универсальной теоретической характеристики степени тесноты связи — индекса корреляции $I_{\eta\cdot\xi}$ (см. формулу (1.6)). Способ построения таких оценок выбирается в зависимости от природы имеющихся у нас выборочных данных и от характера некоторых дополнительных допущений.

Корреляционное отношение. Наиболее привлекательной в этом смысле является ситуация, в которой характер выборочных данных (их количество, «плотность» расположения на плоскости) допускает их группировку по оси объясняющей

переменной и возможность подсчета так называемых «частных» средних ординат $\bar{y}_i$ внутри каждого (i-го) интервала группирования. Пусть такое группирование данных произведено. При этом, как обычно, k — число интервалов группирования по оси абсцисс; m_i ($i = 1, 2, \ldots, k$) — число выборочных точек, попавших в i-й интервал группирования; $\bar{y}_i = (\sum_{j=1}^{m_i} y_{ij})/m_i$ — среднее значение ординат точек, попавших в i-й интервал группирования. Тогда, как легко понять, выборочным аналогом (оценкой) введенной ранее дисперсии σ_f^2 будет величина

$$s^2_{\bar{y}(x)} = \frac{1}{n} \sum_{i=1}^{k} m_i \left(\bar{y}_i - \bar{y}\right)^2, \tag{1.15}$$

где общее среднее $\bar{y} = (\sum_{i=1}^{k} m_i \bar{y}_i)/n$.

Соответственно получаем оценку для $I^2_{\eta \cdot \xi}$ в виде

$$\widehat{\rho}^2_{\eta \cdot \xi} = s^2_{\bar{y}(x)}/s^2_y, \tag{1.16}$$

где выборочная дисперсия s^2_y индивидуальных результатов наблюдения y_{ij} около общего среднего $\bar{y}$ вычисляется по формуле

$$s^2_y = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{m_i} \left(y_{ij} - \bar{y}\right)^2.$$

Величину $\widehat{\rho}_{\eta \cdot \xi}$ принято называть корреляционным отношением зависимой переменной η по независимой переменной ξ. Его вычисление не обременено никакими дополнительными допущениями относительно общего вида регрессионной зависимости (1.1). Однако, в отличие от коэффициента корреляции, корреляционное отношение несимметрично по отношению к исследуемым переменным, т. е., вообще говоря, $\rho_{\eta \cdot \xi} \neq \rho_{\xi \cdot \eta}$. Кроме того, корреляционное отношение, по определению, является величиной неотрицательной[1], так как под ним подразумевается результат извлечения арифметического значения корня квадратного из ρ^2.

В остальном свойства корреляционного отношения во многом похожи на свойства коэффициента корреляции. Из (1.5)

---

[1]Иногда, в частности при *монотонном* характере регрессионной функции (1.1), корреляционному отношению приписывают знак, совпадающий со знаком первой производной этой функции.

и (1.6), в частности, немедленно следует, что подобно коэффициенту корреляции корреляционное отношение не может быть больше единицы.

Из $|\rho| = 1$ следует наличие однозначной функциональной связи между η и ξ, и, наоборот, однозначная функциональная связь между η и ξ свидетельствует о том, что $|\rho| = 1$. Далее, отсутствие корреляционной связи между η и ξ означает, что условные средние $\overline{y}_i$ сохраняют постоянное значение, равное общему среднему $\overline{y}$, а потому $\rho_{\eta/\xi} = 0$. Наоборот, если $\rho_{\eta/\xi} = 0$, то $\overline{y}_i = \overline{y}$, и, следовательно, частные средние $\overline{y}_i$ не зависят от x, т. е. соответствующая линия регрессии параллельна горизонтальной оси.

Отметим, что между $\rho_{\eta/\xi}$ и $\rho_{\xi/\eta}$ нет какой-либо простой зависимости. Некоррелированность η с ξ (т. е. равенство нулю величины $\rho_{\eta/\xi}$) не влечет за собой непосредственно некоррелированности ξ с η. Возможны ситуации, в которых один из этих показателей принимает нулевое значение, в то время как другой равен единице. Допустим, например, что $\eta = \xi^2$ и ξ принимает значения: -1, 0 и $+1$ с вероятностями 1/3 каждое. В этом случае $\rho_{\eta/\xi} = 1$, $\rho_{\xi/\eta} = 0$ (в силу симметрии параболы относительно оси η и симметричности распределения ξ).

Можно показать, что корреляционное отношение ρ не может быть меньше абсолютной величины коэффициента корреляции r, характеризующего зависимость между теми же переменными. В случае линейной зависимости эти две характеристики связи совпадают. Это позволяет использовать величину разности $\widehat{\rho^2_{\eta/\xi}} - \widehat{r^2}$ в качестве меры отклонения регрессионной зависимости от линейного вида (см. п. 6.3.3).

И наконец, все замечания относительно смысловой интерпретации коэффициента корреляции r (в частности, о логическом соотношении понятий «корреляционная зависимость, связь между переменными, их причинная взаимообусловленность») остаются в силе и для корреляционного отношения.

Проверка гипотезы об отсутствии корреляционной связи. Какую величину корреляционного отношения можно признать статистически значимо отличающейся от нуля, т. е. достаточной для статистически обоснованного вывода о наличии корреляционной связи между исследуемыми переменными? Ведь так же, как и в случае прямолинейного типа зависимости, принципиально возможны ситуации, когда отклонение от нуля полученной величины корреляционного отношения $\hat{\rho}$ является статистически незначимым, т. е. обусловленным лишь

неизбежными случайными колебаниями выборки. Для построения соответствующего критерия воспользуемся фактом приближенной $F(k-1, n-k)$-распределенности случайной величины

$$\widehat{F}(0) \quad \frac{\widehat{\rho}^2_{\eta\cdot\xi}}{1-\widehat{\rho}^2_{\eta\cdot\xi}}\cdot\frac{n-k}{k-1},$$

справедливым в предположении, что $I_{\eta\cdot\xi}=0$ (или, что то же, $\rho_{\eta\cdot\xi}=0$) и что условные распределения зависимой переменной $\eta(x)$ при любом фиксированном x описываются нормальным законом с постоянной дисперсией σ^2 (см., например, [65, с. 401]).

Поэтому, если окажется, что

$$\frac{\widehat{\rho}^2_{\eta\cdot\xi}}{1-\widehat{\rho}^2_{\eta\cdot\xi}}\cdot\frac{n-k}{k-1} > v^2_\alpha(k-1, n-k),$$

то гипотеза об отсутствии корреляционной связи между η и ξ отвергается с уровнем значимости α (здесь, как и ранее, $v^2_\alpha(k-1, n-k)$ — 100α %-ная точка F-распределения с числом степеней свободы числителя $k-1$ и знаменателя $n-k$, находится из табл. П.5). При выполнении обратного неравенства значение корреляционного отношения $\widehat{\rho}_{\eta\cdot\xi}$ признается статистически незначимым, т. е. делается вывод об отсутствии корреляционной связи между η и ξ.

Доверительные интервалы для истинного значения корреляционного отношения $\rho_{\eta\cdot\xi}$ можно построить, опираясь на тот факт, что статистика

$$\widehat{F}(\rho)=\frac{\widehat{\rho}^2_{\eta\cdot\xi}}{1-\widehat{\rho}^2_{\eta\cdot\xi}}\cdot\frac{n-k}{k-1} \qquad (1.17)$$

приближенно подчиняется так называемому «нецентральному F-распределению», который оказывается справедливым в предположении $(f(x), \sigma^2)$-нормальности случайных величин $\eta(x)$ и при любом отличном от нуля истинном значении корреляционного отношения $\rho_{\eta\cdot\xi}=\rho$.

Действительно, как известно (см., например, [14, гл. 6]), случайная величина

$$F(\nu_1, \nu_2; a) = \frac{\frac{1}{\nu_1}\sum_{i=1}^{\nu_1}\zeta_i^2}{\frac{1}{\nu_2}\sum_{j=1}^{\nu_2}\gamma_j^2}$$

подчиняется нецентральному F-распределению с числами степеней свободы числителя и знаменателя соответственно ν_1 и ν_2 параметром нецентральности a, если $\zeta_1, \zeta_2, \ldots, \zeta_{\nu_1}, \gamma_1, \gamma_2, \ldots, \gamma_{\nu_2}$ суть взаимно независимые нормальные случайные величины, обладающие одинаковыми дисперсиями, причем $\mathbf{E}\gamma_1 = \mathbf{E}\gamma_2 = \ldots = \mathbf{E}\gamma_{\nu_2} = 0$, а

$$(\mathbf{E}\zeta_1)^2 + (\mathbf{E}\zeta_2)^2 + \ldots + (F\zeta_{\nu_1})^2 = a. \qquad (1.18)$$

Намечая доказательство сформулированного выше утверждения о статистике $\widehat{F}(\rho)$ определенной формулой (1.17), заметим, что в нашем случае в роли случайных величин ζ_i, грубо говоря, выступают значения $\sqrt{m_i}\,(\bar{y}_i - \bar{y})$, а в роли случайных величин $\gamma_j = \gamma_{ij}$ — значения $y_{ij} - \bar{y}_i$. Отметим также следующие соотношения, в справедливости (в некоторых случаях приближенной) которых нетрудно убедиться:

$$\frac{\frac{1}{k-1}\sum_{i=1}^{k} m_i(\bar{y}_i - \bar{y})^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2} = \frac{\frac{1}{k-1}\widehat{\rho}^2_{\eta\cdot\xi}}{\frac{1}{n-k}\left(1 - \widehat{\rho}^2_{\eta\cdot\xi}\right)};$$

$$\mathbf{E}\zeta_i \approx \sqrt{m_i}\,(f(x_i^0) - \bar{f})$$

(здесь $f(x)$ — неизвестная нам функция регрессии η по ξ; x_i^0 — средняя точка i-го интервала группирования по оси абсцисс, а $\bar{f}$ — среднее значение функции регрессии):

$$\mathbf{D}\zeta_i \approx \sigma^2 = \mathbf{D}\eta;$$

$$\mathbf{E}\gamma_{ij} = \mathbf{E}(y_{ij} - \bar{y}_i) = 0;$$

$$\mathbf{D}\gamma_{ii} \approx \sigma^2 = \mathbf{D}\eta.$$

И наконец, параметр нецентральности в соответствии с (1 18) и с учетом (1.6) в нашем случае имеет вид

$$a=\frac{1}{\sigma^2}\sum_{i=1}^{k} m_i\left(f(x_i^0)-\bar{f}\right)^2=\frac{n\sigma_f^2}{\sigma^2}=n\rho^2_{\eta\cdot\xi}.$$

Далее воспользуемся тем (см., например, [30, с. 99]), что распределение статистики $\frac{\nu_1}{\nu_1+a}F(\nu_1,\ \nu_2;\ a)$ при $\nu_1\geqslant 8$ достаточно хорошо аппроксимируется обычным (центральным) F-распределением с числом степеней свободы числителя, приблизительно равным $\nu_1^*=\frac{(\nu_1+a)^2}{(\nu_1+2a)}$, и числом степеней свободы знаменателя, равным ν_2. Поэтому в нашем случае распределение статистики

$$\frac{(n-k)\,\widehat{\rho^2_{\eta\cdot\xi}}}{(k-1)\left(1-\widehat{\rho^2_{\eta\cdot\xi}}\right)}\cdot\frac{k-1}{k-1+n\rho^2_{\eta\cdot\xi}}$$

приближенно описывается F-распределением с числом степеней свободы числителя

$$\nu_1^*=\frac{\left(k-1+n\widehat{\rho^2_{\eta\cdot\xi}}\right)^2}{k-1+2n\widehat{\rho^2_{\eta\cdot\xi}}}\tag{1.19}$$

и числом степеней свободы знаменателя $\nu_2=n-k$.

Таким образом, получаем следующее правило построения приближенных доверительных интервалов для истинного значения корреляционного отношения $\rho_{\eta\cdot\xi}$:

1) пользуясь формулой(1.16), вычисляем точечную оценку $\widehat{\rho^2_{\eta\cdot\xi}}$ для истинного значения корреляционного отношения $\rho^2_{\eta\cdot\xi}$;

2) по формуле (1.19) подсчитываем вспомогательное число степеней свободы ν_1^* числителя для аппроксимирующего центрального F-распределения;

3) задавшись уровнем доверия $P=1-2\alpha$, с помощью табл. П.5 находим $100\,(1-\alpha)\%$-ную точку $v^2_{1-\alpha}(\nu_1^*,\ n-k)$ и $100\alpha\ \%$-ную точку $v^2_{\alpha}(\nu_1^*,\ n-k)$ F-распределения с числом степеней свободы числителя ν_1^* и знаменателя $n-k$;

4) утверждаем, что приблизительно с вероятностью $P=$

$= 1 - 2\alpha$ истинное значение корреляционного отношения $\rho_{\eta\cdot\xi}$ удовлетворяет неравенствам

$$\frac{(n-k)\,\widehat{\rho^2_{\eta\cdot\xi}}}{n\left(1-\widehat{\rho^2_{\eta\cdot\xi}}\right)v^2_{\alpha}} - \frac{k-1}{n} < \rho^2_{\eta\cdot\xi} < \frac{(n-k)\,\widehat{\rho^2_{\eta\cdot\xi}}}{n\left(1-\widehat{\rho^2_{\eta\cdot\xi}}\right)v^2_{1-\alpha}} - \frac{k-1}{n} \qquad (1.20)$$

Проиллюстрируем работоспособность описанного метода на следующем примере. Пусть в результате обработки 132 экспериментальных точек (x_i, y_i) $(i = 1, 2, \ldots, 132)$ получено выборочное значение корреляционного отношения $\hat{\rho} = 0{,}60$. При этом мы воспользовались разбиением диапазона изменения независимой переменной на $k = 12$ равных интервалов группирования. Соответственно получаем в качестве вспомогательного числа степеней свободы числителя величину $\nu_1^* = \frac{(12-1+132\cdot 0{,}36)^2}{12-1+2\cdot 132\cdot 0{,}36} \approx 27$ (частное округляем до целого числа). Задавшись доверительной вероятностью $P = 0{,}90$, из табл. П.5 находим (полагая $\alpha = 0{,}05$):

$$v^2_{0{,}05}(27{,}120) = 1{,}58;$$

$$v^2_{0{,}95}(27{,}120) = \frac{1}{v^2_{0{,}05}(120{,}27)} = \frac{1}{1{,}73} \approx 0{,}58.$$

И наконец, в соответствии с формулой (1.20) находим левый $(\widehat{\rho^2_{\min}})$ и правый $(\widehat{\rho^2_{\max}})$ концы доверительного интервала для истинного значения $\rho^2_{\eta\cdot\xi}$:

$$\widehat{\rho^2_{\min}} = \frac{120\cdot 0{,}36}{132\cdot 0{,}64\cdot 1{,}58} - \frac{11}{132} = 0{,}24;$$

$$\widehat{\rho^2_{\max}} = \frac{120\cdot 0{,}36}{132\cdot 0{,}64\cdot 0{,}58} - \frac{11}{132} = 0{,}87.$$

Таким образом, при точечной оценке $\widehat{\rho_{\eta\cdot\xi}} = 0{,}6$ истинное значение заключено в пределах от $\sqrt{0{,}24}$ до $\sqrt{0{,}87}$ с вероятностью, приблизительно равной 0,9, т. е. $0{,}49 < \rho_{\eta\cdot\xi} < 0{,}93$.

В этом примере хорошо видна *существенная несимметричность концов интервальной оценки* относительно точечной оценки (правый конец интервальной оценки отстоит от точечной оценки на 0,33, в то время как левый конец — всего лишь на 0,11).

Для значений точечных оценок $\widehat{\rho^2}$, близких к нулю или к единице, левый или правый конец интервальной оценки может терять содержательный смысл, выходя за пределы отрезка [0, 1]. В этом случае в качестве левого или правого конца интервальной оценки следует брать соответствующее граничное значение — нуль или единицу (причина подобных нежелательных ситуаций — в аппроксимационном подходе к решению данной задачи). Однако описанный прием все-таки следует признать гораздо более точным, чем применяемый иногда метод построения интервальных оценок для $\rho_{\eta \cdot \xi}$, необоснованно использующий приблизительную $(\rho, \frac{1-\rho^2}{\sqrt{n}})$-нормальность статистики $\widehat{\rho}_{\eta \cdot \xi}$.

Оценка индекса корреляции по несгруппированным данным. Если характер имеющихся у нас выборочных данных (x_i, y_i), $i = 1, 2, \ldots, n$, таков, что не допускает их сколько-нибудь удовлетворительной группировки по оси объясняющей переменной (недостаточно велико n, точки (x_i, y_i) слишком «разрежены» на плоскости), то построению оценок для $I^2_{\eta \cdot \xi}$ мы вынуждены предпослать принятие той или иной гипотезы *об общем виде* регрессионной функции (1.1). О статистических методах проверки подобного рода гипотез см. ниже, гл. 6. Пусть, например, в результате анализа, описанного в гл. 6, нами принята гипотеза о том, что интересующая нас регрессионная зависимость $y_{\text{ср}} = f(x)$ имеет вид алгебраического полинома второго порядка, т. е. $y_{\text{ср}} = \mathbf{E}(\eta | \xi = x) = \theta_0 + \theta_1 x + \theta_2 x^2$. Тогда для оценки введенной ранее характеристики степени тесноты связи между исследуемыми переменными η и ξ — коэффициента детерминации $I^2_{\eta \cdot \xi}$ (или индекса корреляции $I_{\eta \cdot \xi}$) исследователю приходится вначале вычислить оценки $\widehat{\theta}_0$, $\widehat{\theta}_1$ и $\widehat{\theta}_2$ для неизвестных параметров — коэффициентов θ_0, θ_1 и θ_2, входящих в уравнение регрессии (см. гл. 7). И лишь после этого, ориентируясь на правую часть формулы (1.6), мы получим в качестве оценки для $I^2_{\eta \cdot \xi}$ величину:

$$\widehat{I^2_{\eta \cdot \xi}} = 1 - \frac{\frac{1}{n-3} \sum_{i=1}^{n} (y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i - \widehat{\theta}_2 x_i^2)^2}{\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

так как нетрудно показать [65], что величина

$$s_{y(x)}^2 = \frac{1}{n-3} \sum_{i=1}^{n} (y_i - \widehat{\theta}_0 - \widehat{\theta}_1 x_i - \widehat{\theta}_2 x_i^2)^2$$

является в данном случае выборочным аналогом (оценкой) теоретической дисперсии $\overline{\sigma}_{\eta(x)}^2$, участвующей в (1.6).

Пусть в общем случае нами принята гипотеза об общем виде интересующей нас зависимости $y_{ср} = \mathbf{E}(\eta | \xi = x) = f(x; \theta_0, \theta_1, \ldots, \theta_p)$, где f — некоторая известная функция аргумента x, зависящая от $(p+1)$-го неизвестного параметра $\theta_0, \theta_1, \ldots, \theta_p$.

Тогда, пользуясь рекомендациями гл. 7, строим оценки $\widehat{\theta}_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_p$ неизвестных параметров, входящих в описание функции регрессии, после чего вычисляем оценку $\widehat{I_{\eta \cdot \xi}^2}$ коэффициента детерминации $I_{\eta \cdot \xi}^2$ по формуле

$$\widehat{I_{\eta \cdot \xi}^2} = 1 - \frac{\frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - f\left(x_i; \widehat{\theta}_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_p\right)\right)^2}{\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2} \qquad (1.21)$$

З а м е ч а н и е. Можно показать, что, как и следовало ожидать, в частном случае $f(x; \theta_0, \theta_1, \ldots, \theta_p) = \theta_0 + \theta_1 x$ оценка, определяемая соотношением (1.21), совпадает с квадратом выборочного коэффициента корреляции $(\widehat{r^2})$.

Следует отметить, что вычисление и использование выборочных характеристик степени тесноты связи типа (1.21) затруднено по меньшей мере тремя обстоятельствами: 1) необходимостью предварительного выбора общего вида регрессионной зависимости; 2) необходимостью предварительного вычисления оценок для входящих в уравнение регрессии неизвестных параметров; 3) отсутствием строгих рекомендаций по их проверке на статистическую значимость и по построению соответствующих *интервальных* оценок.

1.2. Анализ частных («очищенных») связей

1.2.1. Трудности в интерпретации парных корреляционных характеристик, связанные с опосредованным одновременным влиянием других переменных. Выше (см. п. 1.1.2) приведен пример ситуации, в которой специалисты (технологи) некото-

рое время не могли дать содержательного объяснения статистически выявленной положительной парной корреляционной связи между исследуемыми показателями: процентом брака в трубном производстве и продолжительностью плавки стали, из которой эти трубы делались. Вытекающая отсюда практическая рекомендация — снижать, по возможности, продолжительность плавки с целью понижения процента брака — выглядела явно несостоятельной. И лишь позже выяснилось, что объяснение следует искать в одновременном опосредованном влиянии на эти два показателя *третьего фактора*—типа используемого сырья: использование сырья определенного типа приводило к тенденции одновременного увеличения обоих исследуемых показателей — и длительности плавки, и процента брака. Аналогичные трудности в интерпретации получаемых в результате статистического анализа парных корреляционных характеристик испытывают часто специалисты и в других областях деятельности (см. примеры в п. 1.2.4), причем роль опосредованно влияющего на оба изучаемых показателя фактора может играть и *целое множество* неучтенных переменных.

Это обстоятельство делает необходимым введение таких измерителей статистической связи, которые были бы «*очищены*» от опосредованного влияния других переменных, давали бы оценку степени тесноты интересующей нас связи между переменными y и $x^{(j)}$ (или $x^{(i)}$ и $x^{(j)}$) при условии, что значения остальных переменных зафиксированы на некотором постоянном уровне. В этом случае говорят о статистическом анализе *частных* (или «очищенных») связей и используют соответственно *частные* («*очищенные*») коэффициенты корреляции или другие корреляционные характеристики.

1.2.2. Частные коэффициенты корреляции и их выборочные значения. Поставим в соответствие каждой из ранее введенных парных характеристик статистической связи между переменными $x^{(i)}$ и $x^{(j)}$ ($i, j = 0, 1, \ldots, p$; $x^{(0)} \equiv y$) *частную* («*очищенную*») характеристику, определяемую по той же формуле, но только для *условного* распределения [14, гл. 5] $\varphi(x^{(i)}, x^{(j)} \mid X^{(i,j)} = \mathbf{x})$. Здесь φ — это функция плотности вероятности (если $x^{(i)}$, $x^{(j)}$ непрерывны) или полигон вероятностей (если $x^{(i)}$, $x^{(j)}$ дискретны); $X^{(i,j)}$ — множество переменных, дополняющих пару $(x^{(i)}, x^{(j)})$ до *полного* набора рассматриваемых (наблюдаемых) переменных $X = (x^{(0)}, x^{(1)}, \ldots, x^{(p)})$, а $\mathbf{x}$ — $(p - 1)$-мерный вектор, определяющий заданные уровни, на которых фиксируются значения «мешающих» переменных $X^{(i,j)}$. Есть два взаимосвязанных обстоятельства, которые препятствуют широкому практическому использованию частных характеристик статистической связи в *общем случае*:

частные характеристики статистической связи, вообще говоря, зависят от заданных уровней **x** мешающих переменных (как их выбирать в каждом конкретном случае?);

для подсчета *выборочных* значений частных характеристик статистической связи необходимо иметь выборку *специальной структуры*, обеспечивающей наличие хотя бы нескольких наблюдений при каждом из заданного ряда фиксированных значений **x** мешающих переменных.

Можно, однако, показать (см., например, [20, 65]), что если исследуемые случайные переменные $(x^{(0)}, x^{(1)}, \ldots, x^{(p)})$ подчиняются *многомерному нормальному закону* (см. [14, п. 6.1.5]), то указанные неудобства автоматически исчезают, так как в этом случае частные коэффициенты корреляции не зависят от уровней мешающих переменных **x**, определяющих условие в соответствующем условном распределении. В частности, имеет место следующая формула (при условии невырожденности $(p+1)$-мерного нормального закона):

$$r_{ij.X^{(i,j)}} = \frac{-\mathbf{R}_{ij}}{(\mathbf{R}_{ii}\cdot\mathbf{R}_{jj})^{\frac{1}{2}}}, \tag{1.22}$$

где $r_{ij.X^{(i,j)}}$ — частный коэффициент корреляции между переменными $x^{(i)}$ и $x^{(j)}$ при фиксированных значениях всех остальных переменных $X^{(i,j)}$, а $\mathbf{R}_{kl}$ — алгебраическое дополнение для элемента r_{kl} в определителе корреляционной матрицы $\mathbf{R}$ анализируемых признаков $x^{(0)} \equiv y, x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, т. е. в определителе

$$\det \mathbf{R} = \begin{vmatrix} 1 & r_{01} & r_{01} & \cdots & r_{0p} \\ r_{10} & 1 & r_{12} & \cdots & r_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{p0} & r_{p1} & r_{p2} & \cdots & 1 \end{vmatrix}$$

Формула (1.22), примененная к *трехмерному* признаку $(x^{(0)} \equiv y, x^{(1)}, x^{(2)})$, при $i=0$, $j=1$ и $X^{(i,j)} = x^{(2)}$ дает:

$$r_{01.x^{(2)}} = r_{01(2)} = \frac{r_{01} - r_{02}\cdot r_{12}}{\sqrt{(1-r_{02}^2)(1-r_{12}^2)}}. \tag{1.23}$$

Последовательно присоединяя к мешающим переменным все новые признаки из рассматриваемого набора, можно получить *рекуррентные* соотношения для подсчета частных коэффициентов корреляции $r_{01(2\ldots k+1)}$ порядка k (т. е. при исключении опосредованного влияния k мешающих переменных)

по частным коэффициентам корреляции порядка $k-1 (k = 1, 2, \ldots, p-1)$:

$$r_{01(2,3,\ldots,k+1)} = \frac{r_{01(2\ldots k)} - r_{0k+1(2\ldots k)} \cdot r_{1k+1(2\ldots k)}}{\sqrt{\left(1-r^2_{0k+1(2\ldots k)}\right)\left(1-r^2_{1k+1(2\ldots k)}\right)}}. \qquad (1.23')$$

Выборочные (эмпирические) значения частных коэффициентов корреляции вычисляются по тем же формулам (1.22)—(1.23′) с заменой теоретических значений парных коэффициентов корреляции r_{ij} их выборочными аналогами $\hat{r}_{ij}$ (см. формулу (1.8′)).

Если исследователь имеет дело лишь с тремя-четырьмя переменными ($p = 2, 3$), то удобно пользоваться рекуррентными соотношениями (1.23′). При больших размерностях анализируемого многомерного признака удобнее опираться на формулу (1.22), использующую расчет соответствующих определителей.

Вернемся к общему (негауссовскому) случаю. Практика многомерного статистического анализа показала, что частные коэффициенты корреляции, определенные соотношениями (1.22)—(1.23′), являются, как правило, удовлетворительными измерителями очищенной линейной связи между $x^{(i)}$ и $x^{(j)}$ при фиксированных значениях остальных переменных $X^{(i,j)}$ и в случае, когда распределение анализируемых показателей $(x^{(0)}, x^{(1)}, \ldots, x^{(p)})$ отличается от нормального. *Определив* с помощью формулы (1.22) частный коэффициент корреляции в случае *любого* исходного распределения признаков $(x^{(0)}, x^{(1)}, \ldots, x^{(p)})$, включим его в общий математический инструментарий корреляционного анализа линейных моделей. При этом их можно интерпретировать как показатели тесноты очищенной связи, усредненные по всевозможным значениям фиксируемых на определенных уровнях «мешающих» переменных.

1.2.3. Статистические свойства выборочных частных коэффициентов корреляции (проверка на статистическую значимость их отличия от нуля, доверительные интервалы). При исследовании статистических свойств выборочного частного коэффициента корреляции порядка k (т. е. при исключении опосредованного влияния k мешающих переменных) следует воспользоваться тем (см., например, [20, теорема 4.3.4]), что он распределен точно так же, как и обычный (парный) выборочный коэффициент корреляции между теми же переменными с единственной поправкой: объем выборки надо уменьшить на k единиц, т. е. полагать его равным $n - k$, а не n. Поэтому

при проверке статистически значимого отличия от нуля выборочного частного коэффициента корреляции и при построении для него доверительных интервалов следует пользоваться рекомендациями п. 1.1.3 для парного коэффициента корреляции с заменой n на $n - k$.

1.2.4. Примеры. Рассмотрим некоторые конкретные числовые примеры, демонстрирующие возможный характер искажающего опосредованного влияния «третьих факторов» на корреляцию между двумя анализируемыми переменными.

Пример 1.1. По итогам года 37 однородных предприятий легкой промышленности были зарегистрированы следующие показатели их работы: $x^{(0)} \equiv y$ — среднемесячная характеристика качества ткани (в баллах); $x^{(1)}$ — среднемесячное количество профилактических наладок автоматической линии; $x^{(2)}$ — среднемесячное число обрывов нити.

По матрице исходных данных $(x_i^{(0)}, x_i^{(1)}, x_i^{(2)})_{i=\overline{1,37}}$ были подсчитаны (с помощью (1.8')) выборочные *парные* коэффициенты корреляции $\hat{r}_{ij}$ $(i, j = 0, 1, 2)$: $\hat{r}_{01} = 0{,}105$; $\hat{r}_{02} = 0{,}024$; $\hat{r}_{12} = 0{,}996$.

Проверка «на статистическую значимость», проведенная в соответствии с рекомендациями п. 1.1.3, свидетельствует об отсутствии статистически значимой парной корреляционной связи между качеством ткани, с одной стороны, и числом профилактических наладок и обрывов нити — с другой, что не согласуется с профессиональными представлениями технолога.

Однако расчет *частных* коэффициентов корреляции по формуле (1.23) дает значения $\hat{r}_{01(2)} = 0{,}907$; $\hat{r}_{02(1)} = -0{,}906$, которые вполне соответствуют нашим представлениям о естественном характере связей между изучаемыми показателями.

Построение доверительных интервалов для *истинных* значений $r_{01(2)}$ и $r_{02(1)}$ в соответствии с рекомендациями п. 1.1.3 (в частности, с использованием z-преобразования Фишера, поскольку наш случай характеризуется значениями коэффициентов корреляции, близкими по абсолютной величине к единице) дает: $\operatorname{th} z_1 < r < \operatorname{th} z_2$ с доверительной вероятностью $P = 1 - \alpha$, где $\operatorname{th} z$ — тангенс гиперболический угла z,

$$z_{1,2} = \frac{1}{2} \ln \frac{1 + \hat{r}}{1 - \hat{r}} \mp \frac{u_{\alpha/2}}{\sqrt{(n-1)-3}} - \frac{\hat{r}}{2[(n-1)-1]},$$

а u_q — это q-квантиль стандартного нормального распределения (см. табл. П.3).

В нашем примере $n = 37$, $\alpha = 0{,}05$. Подставляя поочередно в эту формулу значения $\hat{r}_{01(2)} = 0{,}907$ и $\hat{r}_{02(1)} = -0{,}906$ и пользуясь табл. П.7 значений $\operatorname{arcth} \hat{r} = \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}}$, получаем:

$$0{,}821 < r_{01(2)} < 0{,}950;$$

$$-0{,}950 < r_{02(1)} < -0{,}819.$$

Пример 1.2. С целью исследования влияния погодных условий на урожайность кормовых трав Хукер (Journ. Roy. Stat. Soc., 1907, v. 65, p. 1) рассмотрел данные Министерства земледелия Англии за 20 лет, характеризующие урожайность $x^{(0)}$ (в ц/акр), весеннее количество осадков $x^{(1)}$ (в дюймах) и накопленную за весну сумму «активных» (т. е. выше $+5{,}5°$ C) температур $x^{(2)}$ (в градусах по Фаренгейту) однородной в метеорологическом отношении области Англии, включающей в себя группу восточных графств. По выборке $(x_i^{(0)}, x_i^{(1)}, x_i^{(2)})_{i=\overline{1,20}}$ были подсчитаны основные статистические характеристики изучаемой трехмерной случайной величины:

$$\hat{\mathbf{E}} x^{(0)} = 28{,}02; \quad \hat{\mathbf{E}} x^{(1)} = 4{,}91; \quad \hat{\mathbf{E}} x^{(2)} = 594{,}0;$$

$$\hat{\mathbf{D}} x^{(0)} = 19{,}54; \quad \hat{\mathbf{D}} x^{(1)} = 1{,}21; \quad \hat{\mathbf{D}} x^{(2)} = 7225;$$

$$\hat{r}_{01} = 0{,}80; \quad \hat{r}_{02} = -0{,}40; \quad \hat{r}_{12} = -0{,}56.$$

Действительно ли высокая температура в период созревания трав отрицательно влияет на их урожайность (ведь $\hat{r}_{02} = -0{,}40$) или здесь сказывается опосредованное влияние «мешающего» фактора — количества осадков $x^{(1)}$?

Вычисление частных коэффициентов корреляции по рекуррентной формуле (1.23) дает:

$$\hat{r}_{01(2)} = 0{,}759; \quad \hat{r}_{02(1)} = 0{,}097; \quad \hat{r}_{12(0)} = -0{,}436.$$

Как видим, если исключить одновременное влияние количества осадков $x^{(1)}$ на урожайность (с ростом $x^{(1)}$ она повышается) и на сумму активных температур (с ростом $x^{(1)}$ она понижается), то мы уже не обнаружим отрицательной корреляции между температурой и урожайностью ($\hat{r}_{02(1)} = 0{,}097$, в то время как $\hat{r}_{02} = -0{,}40$).

Построение доверительных интервалов для $r_{01(2)}$ и $r_{02(1)}$ (с уровнем доверия $P = 0{,}95$) с использованием z-преобразования Фишера дает в данном случае:

$$0{,}448 < r_{01(2)} < 0{,}890; \qquad -0{,}419 < r_{02(1)} < 0{,}525.$$

Последнее неравенство свидетельствует о том, что у нас нет оснований считать положительную очищенную корреляционную связь между урожайностью и температурой ($r_{02(1)} = 0{,}097$) статистически значимой.

1.3. Анализ множественных связей

1.3.1. Степень тесноты множественной статистической связи и среднеквадратическая ошибка прогноза (аппроксимации) одной переменной по совокупности других. Интуитивно и из смысла рассмотренных выше характеристик степени тесноты статистической связи ясно, что чем теснее эта связь,тем больше информации содержит одна переменная относительно другой, тем точнее можно восстановить (спрогнозировать, аппроксимировать) неизвестное значение одной переменной по заданной величине другой.

При решении практических задач чаще других рассматривается схема, в которой поведение какого-то одного (результирующего) признака η стараются «объяснить» поведением совокупности других (предикторных) переменных $\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})$. Если зафиксировать «значение» $\xi = X$, то из всех возможных способов определения прогнозного (аппроксимирующего) значения $\widehat{y}(X)$ для неизвестного значения $\eta(X)$ наилучшим (в смысле минимума среднего квадрата ошибки прогноза), как оказалось, является *условное среднее значение* анализируемого результирующего показателя η, т. е. величина $f(X) = \mathbf{E}(\eta \mid \xi = X)$, где усреднение производится при условии, что объясняющие переменные зафиксированы на уровне X[1]. Действительно, легко видеть, что для любой другой функции $\tilde{f}(X) \not\equiv f(X)$ будем иметь:

$$\mathbf{E}(\eta - \tilde{f}(X))^2 = \mathbf{E}(\eta - f(X) + f(X) - \tilde{f}(X))^2 =$$
$$= \mathbf{E}[(\eta - f(X))^2 + 2(\eta - f(X))(f(X) - \tilde{f}(X)) + (f(X) - \tilde{f}(X))^2] =$$
$$= \mathbf{E}(\eta - f(X))^2 + 2\mathbf{E}[(\eta - f(X))(f(X) - \tilde{f}(X))] +$$
$$+ \mathbf{E}(f(X) - \tilde{f}(X))^2.$$

---

[1]Если объясняющие переменные $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ *не случайны* по своей природе, то они играют роль обычных параметров, от которых зависит закон распределения случайной величины η.

А поскольку $\mathbf{E}\,[(\eta - f(X))\,(f(X) - \tilde{f}(X))] = \mathbf{E}_X\,\{\mathbf{E}_\eta\,[(\eta - f(X))\,(f(X) - \tilde{f}(X))|X]\} = \mathbf{E}_X\,\{(f(X) - \tilde{f}(X)) \times [\mathbf{E}_\eta(\eta|X) - \mathbf{E}_\eta(f(X)|X)]\} = \mathbf{E}_X\,\{(f(X) - \tilde{f}(X))\,[f(X) - f(X)]\} \equiv 0$ и $\mathbf{E}\,(f(X) - \tilde{f}(X))^2 > 0$, то всегда

$$\mathbf{E}\,(\eta - \tilde{f}(X))^2 > \mathbf{E}\,(\eta - f(X))^2.$$

В этих выкладках использовался способ вычисления математического ожидания *в два этапа*: на первом фиксируются значения X и усреднение производится по значениям η (при фиксированном X), т. е. берется *условное* математическое ожидание при условии, наложенном на ξ; на втором этапе результат усредняется по всевозможным значениям X (нижний индекс у знака математического ожидания показывает, по какчм именно значениям производится усреднение).

Таким образом, мы снова (как и в п. В.5 и 1.1.1) пришли к *функции регрессии* $f(X) = \mathbf{E}\,(\eta|\xi = X)$, на этот раз как к функции от p переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, наиболее точно (в смысле среднеквадратической ошибки) воспроизводящей условное значение исследуемого результирующего показателя $\eta\,(X)$ по заданной величине X объясняющих переменных ξ.

Вернемся теперь к соотношению (1.5), связывающему между собой общую вариацию результирующего показателя ($\sigma_\eta^2 = \mathbf{D}\eta$), вариацию функции регрессии ($\sigma_f^2 = \mathbf{D}f\,(\xi)$) и усредненную (по различным возможным значениям X объясняющих переменных) величину условной дисперсии «регрессионных остатков» ($\overline{\sigma_{\eta\,(X)}^2} = \mathbf{E}_X\mathbf{D}\,[\eta\,|\,\xi = X]$). Оно остается справедливым и в случае *многомерной* предикторной переменной $\xi = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})$ (или $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})$).

Следовательно, так же как и в случае парной зависимости, вариация (случайный разброс) результирующего показателя η складывается из *контролируемой* нами (по значению предикторной переменной X) вариации функции регрессии $f(X)$ и из не поддающегося нашему контролю случайного разброса значений $\eta\,(X)$ (при фиксированном X) относительно функции регрессии $f(X)$. Именно этот неконтролируемый разброс (характеризуемый величиной $\overline{\sigma_{\eta(X)}^2}$) и определяет одновременно и *среднеквадратическую ошибку* прогноза (или аппроксимации) величины результирующего показателя η по значениям предикторных переменных X, и *степень тесноты связи*, существующей между величиной η, с одной стороны, и значениями X — с другой: чем меньше значение $\overline{\sigma_{\eta(X)}^2}$, тем точнее прог-

ноз и тем теснее связь между η и ξ. Эти соображения приводят нас к следующему способу измерения множественной статистической связи.

1.3.2. Множественный коэффициент корреляции и его свойства (общий случай). Опираясь на формулу (1.5), введем измеритель множественной корреляционной связи между η и $(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})$ — множественный *коэффициент корреляции* $R_{\eta\cdot\xi}$ — аналогично тому, как мы определяли в п. 1.1.1 измеритель *парной* связи — индекс корреляции $I_{\eta\cdot\xi}$ (см. формулу (1.6)):

$$R^2_{\eta\cdot\xi} = 1 - \frac{\bar{\sigma}^2_{\eta}(X)}{\sigma^2_{\eta}} \tag{1.24}$$

(квадрат множественного коэффициента корреляции принято называть *коэффициентом детерминации*).

Из соотношения (1.5) немедленно вытекают следующие свойства множественного коэффициента корреляции:

а) $0 \leqslant R_{\eta\cdot\xi} \leqslant 1$;

б) минимальное значение множественного коэффициента корреляции ($R_{\eta\cdot\xi} = 0$) соответствует случаю полного отсутствия корреляционной связи между η и $(\xi^{(1)}, \ldots, \xi^{(p)})$, так как это может быть только при $\sigma^2_f = \mathbf{D}f(\xi) = 0$, т. е. при независимости значений функции регрессии f от величины ее аргументов ξ ($f(\xi) = \text{const}$); это соответствует ситуации, когда усредненная дисперсия «регрессионных остатков» в точности равна общей вариации результирующего показателя;

в) максимальное значение множественного коэффициента корреляции ($R_{\eta\cdot\xi} = 1$) соответствует полному отсутствию варьирования «регрессионных остатков» ($\bar{\sigma}^2_{\eta(X)} = 0$), что означает наличие чисто функциональной связи между η и $(\xi^{(1)}, \ldots, \xi^{(p)})$: $\eta = f(\xi^{(1)}, \ldots, \xi^{(p)})$. Следовательно, в этом случае мы имеем возможность *точно* (*детерминированно*) восстанавливать условные значения $\eta(X) = \{\eta \mid \xi = X\}$ по значениям предикторных переменных X, и соответственно общая вариация результирующего показателя η *полностью объясняется* контролируемой вариацией функции регрессии;

г) *выборочное значение* $\widehat{R}_{\eta\cdot\xi}$ множественного коэффициента корреляции $R_{\eta\cdot\xi}$ определяется на базе системы наблюдений $\{(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; y_i\}_{i=\overline{1,n}}$ по формуле, получающейся из (1.24) заменой участвующих в правой части теоретических характеристик $\bar{\sigma}^2_{\eta(X)}$ и σ^2_{η} их *выборочными аналогами*, т. е.

$$\widehat{R}^2_{\eta\cdot\xi} = 1 - \frac{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - f(x_i^{(1)}, \ldots, x_i^{(p)}; \widehat{\theta}_1, \ldots, \widehat{\theta}_k))^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (1.24')$$

где $f(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; \theta_1, \ldots, \theta_k) = \mathbf{E}(\eta | \xi = X)$ — функция регрессии (η по ξ) *известного* общего вида, зависящая от k параметров $\theta_1, \theta_2, \ldots, \theta_k$, значения которых неизвестны (оцениваются по выборке, см. гл. 6—9)[1], а $\bar{y}$ — выборочное среднее значение результирующего показателя (т. е. $\bar{y} = \sum_{i=1}^{n} y_i/n$);

д) введенные с помощью (1.24) и (1.24′) теоретический и выборочный множественные коэффициенты корреляции формально определены для *любой* ($p+1$)-мерной системы наблюдений. Квадрат их величины $R^2_{\eta\cdot\xi}$ и $\widehat{R}^2_{\eta\cdot\xi}$ показывает, какая доля дисперсии исследуемого результирующего показателя η определяется (*детерминируется*) контролируемой нами вариацией соответствующей функции регрессии $f(X)$. Соответственно оставшаяся доля дисперсии показателя η (т. е. величина $1 - R^2_{\eta\cdot\xi}$ или $1 - \widehat{R}^2_{\eta\cdot\xi}$) объясняется воздействием неконтролируемой случайной остаточной компоненты («регрессионных остатков», «помехи») и определяет ту верхнюю

---

[1]Очевидно, такое определение $\widehat{R}_{\eta\cdot\xi}$ предусматривает априорное знание общего вида функции регрессии $f(X; \Theta)$ и проведение предварительных расчетов по статистическому оцениванию неизвестных значений участвующих в ее записи параметров $\Theta = (\theta_1, \ldots, \theta_k)$. Ниже (см. п. 1.3.3) мы увидим, что последнее неудобство автоматически устраняется при работе с данными из *нормальных* генеральных совокупностей, причем получаемые в этом случае удобные формулы и рекомендации могут быть использованы как *приближенные* и в общем случае. Что касается неудобства, связанного с необходимостью априорного знания общего вида функции регрессии $f(X, \Theta)$, то в зависимости от конкретизации задачи и условий сбора исходных данных могут быть использованы следующие альтернативные подходы: а) предварительное разбиение области значений предикторных переменных X на *гиперпараллелепипеды группирования* Δ_j, вычисление условных средних $\bar{y}_j$ результирующего показателя по наблюдениям, попавшим в Δ_j, и замена дисперсии $\sigma^2_{\eta(X)}$ в формуле (1.24) ее оценкой, построенной по разбросу значений η внутри каждого гиперпараллелепипеда группирования относительно своих условных средних $\bar{y}_j$ (что приводит нас к обобщению понятия *корреляционного отношения* на многомерный случай, ср. с п. 1.1.5); б) использование *непараметрической* и *частично-параметрической* техники оценивания функции регрессии $f(X)$ (см. гл. 10).

границу точности, которой мы можем добиться при восстановлении (прогнозировании, аппроксимации) значения результирующего показателя η по заданным значениям X объясняющих переменных ξ.

1.3.3. Вычисление и свойства множественного коэффициента корреляции в рамках линейных нормальных моделей. Если предположить, что исходные статистические данные $\{(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; y_i\}_{1=\overline{1,n}}$ могут интерпретироваться как выборка объема n из $(p+1)$-мерной *нормальной* генеральной совокупности с вектором средних значений

$$\mathbf{M}=\begin{pmatrix}\mathbf{E}\xi^{(1)}\\ \vdots\\ \mathbf{E}\xi^{(p)}\\ \mathbf{E}\eta\end{pmatrix}=\begin{pmatrix}m^{(1)}\\ \vdots\\ m^{(p)}\\ m^{(0)}\end{pmatrix}=\begin{pmatrix}\mathbf{M}_\xi\\ m^{(0)}\end{pmatrix}$$

и ковариационной матрицей (см. сноску перед формулой (1.3))

$$\Sigma=\begin{pmatrix}\Sigma_{\xi\xi} & \Sigma_{\xi\eta}\\ \Sigma_{\eta\xi} & \Sigma_{\eta\eta}\end{pmatrix},$$

то из (1.3)—(1.4) сразу следует:

а) функция $f(X)$ регрессии η по $\xi=(\xi^{(1)}, \ldots, \xi^{(p)})$ *линейна* по аргументам, а именно:

$$f(X)=\mathbf{E}(\eta|\xi=X)=m^{(0)}+(\sigma_{01}\,\sigma_{02}\,\ldots\,\sigma_{0p})\times$$

$$\times\begin{pmatrix}\sigma^{11} & \sigma^{12} & \ldots & \sigma^{1p}\\ \sigma^{21} & \sigma^{22} & \ldots & \sigma^{2p}\\ \cdot & \cdot & \cdot & \cdot\\ \sigma^{p1} & \sigma^{p2} & \ldots & \sigma^{pp}\end{pmatrix}(X-\mathbf{M}_\xi), \qquad (1.25)$$

где $\sigma_{ij}=\mathbf{E}\,[(\xi^{(i)}-m^{(i)})\,(\xi^{(j)}-m^{(j)})]$ — ковариации анализируемых переменных (мы полагаем, для единообразия записи, $\xi^{(0)}\equiv\eta$), а σ^{ij} — элементы матрицы $\Sigma_{\xi\xi}^{-1}$;

б) условная (остаточная) дисперсия $\sigma^2_{\eta(X)}=\mathbf{D}\,(\eta|\xi=X)$ результирующего показателя η не зависит от того, на каких уровнях X фиксируются значения объясняющих переменных ξ, в частности

$$\sigma^2_{\eta\,(X)}=\mathbf{D}(\eta|\xi=X)=\sigma^2_\eta\cdot(1-R^2_{\eta\cdot\xi}). \qquad (1.26)$$

Условимся относить подобные ситуации к *первому типу линейных нормальных моделей.*

Разрешая выражение (1.26) относительно $R^2_{\eta\cdot\xi}$, мы приходим (с учетом постоянства по X величины $\sigma^2_{\eta(X)}$ в данном случае) к ранее введенному определению множественного коэффициента корреляции (1.24).

Отнесем ко *второму типу линейных нормальных моделей* тот частный случай «схемы B» (т. е. зависимости случайного результирующего показателя η от неслучайных объясняющих переменных X, см. § В.5), в котором функция регрессии $f(X)$ линейна по X, а остаточная случайная компонента $\varepsilon(X)$ подчиняется нормальному закону с постоянной (не зависящей от X) дисперсией σ^2_ε. В этом случае линейность регрессии, гомоскедастичность (постоянство условной дисперсии $\sigma^2_{\eta(X)} = \sigma^2_\varepsilon$) и формула (1.26) следуют непосредственно из определения модели и из (1.24).

Можно показать (см. например, [65, гл. 27]), что при статистической обработке выборок, извлеченных из линейно-нормальных генеральных совокупностей, множественный коэффициент корреляции $R_{\eta\cdot\xi}$ и его выборочное значение $\widehat{R}_{\eta\cdot\xi}$ обладают рядом дополнительных свойств (приведенные ниже формулы и свойства теоретического множественного коэффициента корреляции $R_{\eta\cdot\xi}$ автоматически переносятся на выборочный $\widehat{R}_{\eta\cdot\xi}$ заменой участвующих в них теоретических характеристик соответствующими выборочными значениями).

1. *Вычисление $R_{\eta\cdot\xi}$ по матрице парных коэффициентов корреляции.* Обозначая, как и прежде, $(p+1)\times(p+1)$-корреляционную матрицу $(r_{ij})_{ij=0,1\ldots,p}$ через $\mathbf{R}$, а алгебраическое дополнение элемента r_{kl} в ее определителе через $|\mathbf{R}|_{kl}$, имеем

$$R^2_{\eta\cdot\xi} = 1 - \frac{\det \mathbf{R}}{|\mathbf{R}|_{00}}. \tag{1.27}$$

2. *Вычисление $R_{\eta\cdot\xi}$ по частным коэффициентам корреляции*

$$R^2_{\eta\cdot\xi} = 1-(1-r^2_{01})(1-r^2_{02\,(1)})(1-r^2_{03\,(12)})\ldots$$
$$\ldots(1-r^2_{0p\,(12\ldots p-1)}). \tag{1.28}$$

3. *Множественный коэффициент корреляции мажорирует любой парный или частный коэффициент корреляции, характеризующий статистическую связь результирующего показателя, т. е.*

$$R_{\eta\cdot\xi} \geqslant |r_{0j(I_j)}|, \tag{1.29}$$

где $j = 1, 2, \ldots, p$, а I_j — любое подмножество множества индексов $I_0 = \{1, 2, \ldots, p\}$, не содержащее индекса j (соотношение (1.29) следует из (1.28)). Напоминаем, что $\xi^{(0)} \equiv \eta$.

4. *Присоединение каждой новой предсказывающей переменной не может уменьшить величины* R (независимо от порядка присоединения), т. е.

$$R_{\eta \cdot \xi^{(1)}} \leqslant R_{\eta \cdot (\xi^{(1)}, \xi^{(2)})} \leqslant R_{\eta \cdot (\xi^{(1)}, \xi^{(2)}, \xi^{(3)})} \leqslant \ldots$$
$$\ldots \leqslant R_{\eta \cdot (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})}. \tag{1.30}$$

5. *Множественный коэффициент корреляции* $R_{\eta \cdot \xi}$ *может быть определен как максимальное значение обычного парного коэффициента корреляции между* η *и линейной комбинацией* $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}$ (*максимум — по всевозможным линейным комбинациям*) *либо как обычный парный коэффициент корреляции между* η *и условным математическим ожиданием* $\mathbf{E}(\eta|X)$.

6. *Статистические свойства выборочного множественного коэффициента корреляции* $\widehat{R}_{\eta \cdot \xi}$ (*распределение, моменты, доверительные интервалы*) состоят в следующем.

Для проверки гипотезы H_0: $R_{\eta \cdot \xi} = 0$, т. е. для выяснения вопроса, можно ли считать выборочное значение множественного коэффициента корреляции $\widehat{R}_{\eta \cdot \xi}$ *статистически значимо отличающимся от нуля*, пользуются фактом $F(p, n-p-1)$-распределенности случайной величины

$$\widehat{F}(\widehat{R}) = \frac{\widehat{R}^2_{\eta \cdot \xi}}{1 - \widehat{R}^2_{\eta \cdot \xi}} \cdot \frac{n-p-1}{p},$$

справедливым в рамках обоих рассмотренных выше типов линейно-нормальных моделей при условии, что истинное значение множественного коэффициента корреляции $R_{\eta \cdot \xi}$ равно нулю. Если окажется, что $\widehat{F}(\widehat{R}) > v^2_{\alpha}(p, n-p-1)$, то гипотеза об отсутствии множественной корреляционной связи между η и $(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})$ отвергается при уровне значимости критерия, равном α (здесь, как и ранее, $v^2_{\alpha}(p, n-p-1)$ — $100\alpha\%$-ная точка F-распределения с числом степеней свободы числителя p и знаменателя $n-p-1$ находится из табл. П.5).

Можно показать (см. [65, гл. 27]), что в условиях *второго типа* линейно-нормальных моделей (объясняющие переменные X неслучайны) описанный критерий является равномерно наиболее мощным. Это вытекает из того, что при $R_{\eta \cdot \xi} \neq 0$ величина $\widehat{F}(\widehat{R})$ *подчинена нецентральному* $F(p, n-p-$

-1; $nR^2_{\eta\cdot\xi}$) *-распределению* с параметром нецентральности, равным $nR^2_{\eta\cdot\xi}$.

Последним обстоятельством можно воспользоваться и при приближенном построении доверительных интервалов для неизвестного истинного значения $R^2_{\eta\cdot\xi}$. В точности повторяя рассуждения п. 1.1.5, относящиеся к построению доверительных интервалов для неизвестной величины квадрата корреляционного отношения $\rho^2_{\eta\cdot\xi}$ (см. формулы (1.17)—(1.20)), мы придем к следующей рекомендации по построению интервальной оценки для $R^2_{\eta\cdot\xi}$*, *справедливой, правда, лишь при* $p \geqslant 8$:

с доверительной вероятностью, приблизительно равной $1-2\alpha$ (величина α задана), выполняется неравенство

$$\widehat{R}^2_{\eta\cdot\xi}\cdot\frac{1-\frac{p+1}{n}}{v^2_{\alpha}(\nu_1,\nu_2)\cdot\left(1-\widehat{R}^2_{\eta\cdot\xi}\right)}-\frac{p}{n}<R^2_{\eta\cdot\xi}<\widehat{R}^2_{\eta\cdot\xi}\times$$

$$\times\frac{1-\frac{p+1}{n}}{v^2_{1-\alpha}(\nu_1,\nu_2)\cdot\left(1-\widehat{R}^2_{\eta\cdot\xi}\right)}-\frac{p}{n}, \qquad (1.31)$$

в котором $v^2_q(\nu_1, \nu_2)$ — $100q$%-ная точка центрального F-распределения с числом степеней свободы числителя

$$\nu_1=\left[\frac{\left(p+n\widehat{R}^2_{\eta\cdot\xi}\right)^2}{p+2n\widehat{R}^2_{\eta\cdot\xi}}\right] \qquad (1.32)$$

и знаменателя $\nu_2 = n-p-1$ (в (1.32) символ $[a]$ обозначает ближайшее целое число к a).

Однако в условиях *первого типа* линейно-нормальных моделей (наблюдения $(x^{(1)}_i, x^{(2)}_i, \ldots, x^{(p)}_i, y_i)$ извлечены из $(p+1)$-мерной нормальной генеральной совокупности; соответственно объясняющие переменные $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}$ — *случайные* величины) распределение величины $\widehat{R}^2_{\eta\cdot\xi}$ при $R_{\eta\cdot\xi} \neq 0$ и конечных объемах выборки (n) существенно отличается от того распределения $\widehat{R}^2_{\eta\cdot\xi}$, которое мы имели при неслучайных объясняющих переменных (можно, правда, показать, что при $n \to \infty$ распределение случайной величины $n\widehat{R}^2_{\eta\cdot\xi}$ сходится в линейно-нормальных моделях и первого и

---

*Для построения доверительных интервалов для неизвестного истинного значения множественного коэффициента корреляции $R_{\eta\cdot\xi}$ читатель может воспользоваться также специальными номограммами, приведенными в [50] для случаев $p = 3, 5, 7$.

второго типа к нецентральному χ^2-распределению с числом степеней свободы, равным p, и с параметром нецентральности, равным nR^2). Р. Фишер [183] и ряд других исследователей занимались изучением распределения величины $\widehat{R}_{\eta\cdot\xi}$ в условиях первого типа линейно-нормальных моделей (различные представления соответствующей функции плотности вероятности можно найти, например, в [65, гл. 27]).

Приведем здесь лишь выражения для первых двух моментов интересующей нас величины.

Случай $R_{\eta\cdot\xi}=0$.

$$\mathbf{E}\widehat{R}^2_{\eta\cdot\xi}=\frac{p}{n-1}; \tag{1.33}$$

$$\mathbf{D}\widehat{R}^2_{\eta\cdot\xi}=\frac{2(n-p-1)p}{(n^2-1)(n-1)}\approx\frac{2p}{n^2}. \tag{1.34}$$

Случай $R_{\eta\cdot\xi}\neq 0$.

$$\mathbf{E}\widehat{R}^2_{\eta\cdot\xi}=R^2_{\eta\cdot\xi}+\frac{p}{n-1}(1-R^2_{\eta\cdot\xi})-\frac{2(n-p)}{n^2-1}R^2_{\eta\cdot\xi}(1-R^2_{\eta\cdot\xi})+$$

$$+0\left(\frac{1}{n^2}\right)\approx R^2_{\eta\cdot\xi}+\frac{p}{n-1}(1-R^2_{\eta\cdot\xi}); \tag{1.33'}$$

$$\mathbf{D}\widehat{R}^2_{\eta\cdot\xi}=\frac{2(1-R^2_{\eta\cdot\xi})^2(n-p-1)p}{(n^2-1)(n-1)}+0\left(\frac{1}{n^2}\right)\approx$$

$$\approx\frac{2p(n-p-1)}{(n-1)(n^2-1)}(1-R^2_{\eta\cdot\xi})^2. \tag{1.34'}$$

Скорректированная (на несмещенность) оценка $R^2_{\eta\cdot\xi}$. По формулам (1.33), (1.33′) мы видим, что при вычислении выборочных значений $\widehat{R}^2_{\eta\cdot\xi}$ в соответствии с рекомендациями (1.27), (1.28), относящимися к условиям линейно-нормальных моделей, получаются *смещенные* (а при ограниченных объемах выборок n и большом числе p предсказывающих переменных — *существенно смещенные*) оценки для неизвестного истинного значения $R^2_{\eta\cdot\xi}$. Поэтому желательно попытаться перейти к некоторой другой оценке $\widehat{R^*}^2_{\eta\cdot\xi}$ неизвестного теоретического значения $R^2_{\eta\cdot\xi}$ путем такой коррекции оценки $\widehat{R}^2_{\eta\cdot\xi}$, которая позволила бы устранить это смещение.

В [233] показано, что несмещенной оценкой коэффициента $R^2_{\eta\cdot\xi}$ служит статистика

$$\widehat{R}^{*2}_{\eta\cdot\xi} = 1 - \frac{n-3}{n-p-1}(1-\widehat{R}^2_{\eta\cdot\xi})\cdot F\left(1;\ 1;\ \frac{1}{2}(n-p+1);\ 1-R^2_{\eta\cdot\xi}\right), \qquad (1.35)$$

где $2 \leqslant p < n-1$, а $F(a;\ b;\ c;\ d)$ — *гипергеометрическая функция* (см., например, [1, с. 370]).

Простая аппроксимация правой части (1.35) дает:

$$\widehat{R}^{*2}_{\eta\cdot\xi} \approx 1-(1-\widehat{R}^2_{\eta\cdot\xi})\frac{n-1}{n-p-1}. \qquad (1.35')$$

Из последней формулы видно, что «подправленная» оценка $\widehat{R}^{*2}_{\eta\cdot\xi}$ всегда меньше смещенной оценки $\widehat{R}^2_{\eta\cdot\xi}$.

Отметим, что при малых истинных значениях $R^2_{\eta\cdot\xi}$ и при «не слишком малых» величинах отношения p/n подправленные оценки, подсчитанные по формулам (1.35) и (1.35′), могут принимать *отрицательные* значения. Можно устранить абсурдность отрицательных значений оценки, используя в качестве «еще раз подправленной» оценки величину

$$\widehat{R}^{**2}_{\eta\cdot\xi} = \max(\widehat{R}^{*2}_{\eta\cdot\xi},\ 0)$$

(правда, $\widehat{R}^{**2}_{\eta\cdot\xi}$ уже не будет несмещенной оценкой).

1.3.4. Примеры. Вернемся к ранее рассмотренным примерам и оценим в них степень тесноты множественной связи между результирующим показателем, с одной стороны, и набором объясняющих переменных — с другой. Будем пользоваться рекомендациями (а именно формулами (1.27), (1.28)), правомерность которых строго обоснована лишь в рамках линейно-нормальных моделей.

Пример 1.1. Оценка $\widehat{R}_{y\cdot(x^{(1)}x^{(2)})}$ коэффициента множественной корреляции между характеристикой качества ткани y и совокупностью двух факторов: количеством профилактических наладок $x^{(1)}$ и числом обрывов нити $x^{(2)}$, подсчитанная с помощью формулы (1.28), дает:

$$\widehat{R}^2_{y\cdot(x^{(1)}x^{(2)})} = 1-(1-\widehat{r}^2_{01})(1-\widehat{r}^2_{02(1)}) =$$
$$= 1-[1-(0{,}105)^2][1-(0{,}906)^2] =$$
$$= 1-0{,}989\cdot 0{,}179 = 1-0{,}177 = 0{,}823.$$

Отсюда $\widehat{R}_{y\cdot(x^{(1)}x^{(2)})} = \sqrt{0{,}823} = 0{,}9072$.

В данном примере мы не можем воспользоваться формулами (1.31)—(1.32) для построения доверительного интервала для $R^2_{y.(x^{(1)}x^{(2)})}$, поскольку они дают удовлетворительную точность лишь при $p \geqslant 8$.

Пример 1.2. Оценка $\hat{R}_{y \cdot x^{(1)}x^{(2)}}$ коэффициента множественной корреляции между урожайностью кормовых трав ($y \equiv x^{(0)}$) и природными факторами — весенним количеством осадков ($x^{(1)}$) и накопленной суммой «активных» температур ($x^{(2)}$), подсчитанная по формуле (1.28), дает:

$$\hat{R}^2_{y \cdot (x^{(1)}\, x^{(2)})} = 1 - (1 - \hat{r}^2_{01})(1 - \hat{r}^2_{02(1)}) =$$

$$= 1 - [1 - (0{,}80)^2][1 - (1{,}097)^2] =$$

$$= 1 - 0{,}36 \cdot 0{,}99 = 0{,}6436.$$

Отсюда $\hat{R}_{y \cdot (x^{(1)}\, x^{(2)})} = \sqrt{0{,}6436} = 0{,}802.$

ВЫВОДЫ

1. Приступая к статистическому исследованию зависимостей между анализируемыми переменными, исследователь должен в первую очередь установить *сам факт наличия* статистических связей и попытаться *измерить степень их тесноты*. В качестве основных измерителей степени тесноты связей между количественными переменными в практике статистических исследований используются: индекс корреляции, корреляционное отношение, парные, частные и множественные коэффициенты корреляции, коэффициент детерминации.

2. *Парные* корреляционные характеристики позволяют измерять степень тесноты статистической связи между парой переменных *без учета опосредованного или совместного влияния других показателей*. Вычисляются (оцениваются) они по результатам наблюдений только анализируемой пары показателей.

3. Факт установления тесной статистической связи между переменными не является, вообще говоря, достаточным основанием для доказательства существования *причинно-следственной связи между этими переменными*.

4. Парные и частные *коэффициенты корреляции* являются измерителями степени тесноты линейной связи между переменными. В этом случае корреляционные характеристики могут оказаться как положительными, так и отрицательными в за-

висимости от одинаковой или противоположной тенденции взаимосвязанного изменения анализируемых переменных. При положительных значениях коэффициента корреляции говорят о наличии *положительной линейной статистической связи*, при отрицательных — об *отрицательной*.

5. При наложении случайных ошибок на значения исследуемой пары переменных (например, ошибок измерения) оценка статистической связи между исходными переменными, построенная по наблюдениям, оказывается искаженной. В частности, получаемые при этом оценки коэффициентов корреляции будут *заниженными*. Существуют методы, позволяющие учесть это искажение.

6. Измерителем степени тесноты связи *любой формы* является *корреляционное отношение*, для вычисления которого необходимо разбить область значений предсказывающей переменной X на интервалы (гиперпараллелепипеды) группирования. Возможна параметрическая модификация корреляционного отношения, при которой вычисление соответствующих выборочных значений не требует предварительного разбиения на интервалы группирования.

7. *Частный коэффициент корреляции* позволяет оценить степень тесноты линейной связи между двумя переменными, *очищенной от опосредованного влияния* других факторов. Для его расчета необходима исходная информация как по анализируемой паре переменных, так и по всем тем переменным, опосредованное («мешающее») влияние которых мы хотим элиминировать.

8. *Множественный* (*совокупный*) *коэффициент корреляции* измеряет степень тесноты статистической связи (любой формы) между некоторым (результирующим) показателем, с одной стороны, и совокупностью других (объясняющих) переменных — с другой. Формально он определен для любой многомерной системы наблюдений. Квадрат его величины (называемый *коэффициентом детерминации*) показывает, какая доля дисперсии исследуемого результирующего показателя определяется (детерминируется) совокупным влиянием контролируемых нами (в виде функции регрессии) объясняющих переменных. Оставшаяся «необъясненной» доля дисперсии результирующего показателя определяет ту верхнюю границу точности, которой мы можем добиться при восстановлении (прогнозировании, аппроксимации) значения результирующего показателя по заданным значениям объясняющих переменных.

9. Наиболее удобные свойства (рекомендации по вычислению, по интерпретации, статистические свойства) выборочный ко-

эффициент корреляции имеет в рамках линейно-нормальных моделей, т. е. в одном из двух типов ситуаций:

а) обрабатываемые статистические данные $\{(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; y_i\}_{i=\overline{1,n}}$ образуют выборку из $(p+1)$-мерной нормальной генеральной совокупности;

б) результирующий показатель η связан с объясняющими переменными $(x^{(1)}, \ldots, x^{(p)})$ линейной регрессионной зависимостью типа B (см. § B.5), причем остаточная случайная компонента подчиняется нормальному закону с постоянной (не зависящей от $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$) дисперсией. В этом случае разработаны рекомендации по проверке выборочного множественного коэффициента корреляции на его статистически значимое отличие от нуля, по построению доверительных интервалов для неизвестного истинного значения множественного коэффициента корреляции.

Глава 2. АНАЛИЗ СТАТИСТИЧЕСКОЙ СВЯЗИ МЕЖДУ ПОРЯДКОВЫМИ (ОРДИНАЛЬНЫМИ) ПЕРЕМЕННЫМИ

Напомним (см. [14, § 5.3, 10.2]), что *порядковая* (*ординальная*) переменная позволяет *упорядочивать* статистически обследованные объекты по степени проявления в них анализируемого свойства. Исследователь обращается к порядковым переменным в ситуациях, когда шкала непосредственного *количественного* измерения степени проявления этого свойства в объекте ему не известна (в том числе по причине объективного отсутствия таковой) или имеет условный смысл и интересует его только как *вспомогательное средство для последующего ранжирования рассматриваемых объектов*. К подобным ситуациям относится рассмотрение таких переменных, как «интегральный (сводный) показатель эффективности функционирования социально-экономической системы» (специалиста, предприятия, научно-производственного объединения и т. п.), «качество (мера оптимальности) структуры потребительского бюджета семьи», «качество жилищных условий семьи», «степень прогрессивности предлагаемого проекта решения социально-экономической, технической или другой проблемы» и т. п.

Таким образом, в отличие от статистического анализа k-го ($k = 0, 1, 2, \ldots, p$) *количественного* признака $x^{(k)}$, когда в результате его измерения (наблюдения) на объектах мы могли каждому статистически обследованному объекту O_i по-

ставить в соответствие некоторую, *измеренную в физически интерпретируемой* шкале числовую характеристику $x_i^{(k)}$, результатом измерения *порядковой* переменной является приписывание каждому из обследованных объектов некоторой *условной числовой метки*, обозначающей место этого объекта в ряду из всех n анализируемых объектов, упорядоченном по убыванию степени проявления в них k-го изучаемого свойства. В этом случае $x_i^{(k)}$ называют *рангом* i-го объекта по k-му признаку.

В зависимости от типа изучаемой ситуации (1) шкала измерения признака $x^{(k)}$ не известна исследователю или отсутствует вовсе; 2) существуют косвенные или частные количественные показатели, в соответствии со значениями которых можно определять место каждого объекта O_i в ряду, упорядоченном по анализируемому свойству $x^{(k)}$) сам процесс упорядочения объектов O_1, O_2, ..., O_n производится либо с использованием *экспертной информации*, т. е. с привлечением экспертов, либо *формализованно* — путем перехода от исходного ряда наблюдений вспомогательного (косвенного, частного) количественного признака к соответствующему *вариационному ряду* [14, п. 5.6.4].

2.1. Ранговая корреляция

2.1.1. Исходные статистические данные (таблица или матрица рангов типа «объект — свойство»). Итак, в результате измерения $p+1$ порядковых переменных $x^{(0)} \equiv y$, $x^{(1)}$, ..., $x^{(p)}$ на каждом из n анализируемых объектов O_1, O_2, ..., O_n мы получаем таблицу (матрицу) исходных данных следующего вида (табл. 2.1).

В этой таблице элемент $x_i^{(k)}$ задает порядковое место (*ранг*), которое занимает объект O_i в ряду всех статистически обследованных объектов, упорядоченном по убыванию степени проявления k-го анализируемого свойства (т. е. по переменной $x^{(k)}$).

Очевидно, если рассмотреть *столбец* с номером k этой таблицы ($k = 0, 1, ..., p$), то он будет представлять перестановку из n элементов, а именно перестановку из n натуральных чисел 1, 2, ..., n, определяющую порядковые места объектов O_1, O_2, ..., O_n в ряду, упорядоченном по свойству $x^{(k)}$.

Замечание о случаях неразличимости рангов (*«объединенные ранги»*). При упорядочении объектов по какому-либо свойству $x^{(k)}$ ($k = 0, 1, ..., p$) могут встретиться ситуации, когда два объекта или целая группа их оказываются неразличимы-

Таблица 2.1

Порядковый номер объекта («объект»)	Порядковый номер исследуемой переменной («свойство»)						
	0	1	3	...	k	...	p
1	$x_1^{(0)}$	$x_1^{(1)}$	$x_1^{(2)}$	...	$x_1^{(k)}$	...	$x_1^{(p)}$
2	$x_2^{(0)}$	$x_2^{(1)}$	$x_2^{(2)}$	...	$x_2^{(k)}$	...	$x_2^{(p)}$
. . .	...	...	...	...	...	...	...
i	$x_i^{(0)}$	$x_i^{(1)}$	$x_i^{(2)}$	...	$x_i^{(k)}$	...	$x_i^{(p)}$
...	...	...	...	...	...	...	...
n	$x_n^{(0)}$	$x_n^{(1)}$	$x_n^{(2)}$	...	$x_n^{(k)}$	...	$x_n^{(p)}$

ми с точки зрения степени проявления в них этого свойства. Тогда каждому из объектов этой однородной группы приписывается ранг, равный среднему арифметическому значению тех мест, которые они делят, а полученные таким образом ранги принято называть *«объединенными»* (или «связными»). Так, например, упорядочивая семь альтернативных проектов A, B, C, D, E, F, G перспективного развития некоторой подотрасли с точки зрения их народнохозяйственной эффективности, эксперт поставил на 1-е место проект C, на 2-е — проект A, далее располагал проекты B, D и E, которые считал неразличимыми (равноценными) по эффективности, а последнее место отвел проектам F и G. Тогда соответствующий столбец таблицы «объект—свойство» будет состоять из следующих компонент:

$$x_A = 2;\quad x_B = \frac{3+4+5}{3} = 4;\quad x_C = 1;\quad x_D = x_E = \frac{3+4+5}{3} = 4;$$

$$x_F = x_G = \frac{6+7}{2} = 6{,}5.$$

Мы видим, что появление объединенных рангов может привести к *дробным значениям* рангов, составляющих массив исходных статистических данных (значения рангов, соответствующие 6-му и 7-му проектам). При отсутствии объединенных рангов область возможных значений переменных $x^{(k)}$, очевидно, ограничивается множеством первых n чисел натурального ряда, где n — число сравниваемых объектов.

Мы увидим далее, что наличие объединенных рангов несколько усложняет вычислительные процедуры, связанные со статистическим анализом соответствующих корреляционных характеристик.

2.1.2. Понятие ранговой корреляции. Под *ранговой корреляцией* понимается статистическая связь между порядковыми переменными. В статистической практике эта связь анализируется на основании исходных статистических данных, представленных упорядочениями (ранжировками) n рассматриваемых объектов по разным свойствам (см. столбцы табл. 2.1). Есть ли хоть какая-то согласованность (или связь) между упорядочением анализируемых объектов по свойству $x^{(k)}$ и упорядочением тех же объектов по другому свойству $x^{(j)}$? Можно ли измерить и проанализировать совокупную статистическую связь, существующую между ранжировками одних и тех же объектов $O_1, O_2, \ldots, O_n$, полученными в соответствии со степенью проявления в них сначала свойства $x^{(k_1)}$ (1-й способ упорядочения), затем — свойства $x^{(k_2)}$ (2-й способ упорядочения)? Таким образом, речь идет о системе понятий и методов, позволяющих измерять и анализировать статистическую связь, существующую между двумя или несколькими ранжировками одного и того же конечного множества объектов $O_1, O_2, \ldots, O_n$.

Система этих понятий и методов и составляет раздел математической статистики, который принято называть анализом *ранговых корреляций*. Методы ранговой корреляции широко используются, в частности, при организации и статистической обработке различного рода систем экспертных обследований (см., например, [126, 131]).

2.1.3. Основные задачи статистического анализа связеи между ранжировками. Предположим, мы ввели измерители парной и множественной ранговой статистической связи (см. ниже п. 2.2—2.3). Тогда, опираясь на эти характеристики, исследователь чаще всего пытается решить следующие три основные задачи статистического анализа структуры и характера связей, существующих между изучаемыми порядковыми переменными.

Задача А: *анализ структуры имеющейся совокупности упорядочений* $X^{(k)}_{\cdot} = (x^{(k)}_1, x^{(k)}_2, \ldots, x^{(k)}_n)'$, $k = 0, 1, \ldots, p$. Интерпретируя каждое упорядочение $X^{(k)}_{\cdot}$ как точку в n-мерном пространстве, можно представить, например, три наиболее характерных типа такой структуры: 1) анализируемые точки *равномерно* разбросаны по всей области своих возможных значений (определяемой неравенствами $1 \leqslant x^{(k)}_i \leqslant$

$\leqslant n$, $i = 1, 2, \ldots, n$), что означает отсутствие какой-либо связи или согласованности в представляемых ими ранжировках; 2) расположение $p + 1$ точек таково, что часть из них образует ядро из близко лежащих друг от друга точек («сгусток»), а остальные произвольно разбросаны относительно этого ядра. В этом случае существование ядра обеспечивает наличие подмножества согласованных переменных; 3) анализируемые точки — ранжировки располагаются в пространстве несколькими относительно далеко отстоящими друг от друга ядрами («сгустками»), что означает наличие нескольких подмножеств переменных таких, что переменные внутри одного подмножества обнаруживают высокую статистическую взаимосвязь, тогда как согласованности между переменными, взятыми из разных таких подсовокупностей, практически не существует.

Задача В: *анализ интегральной (совокупной) согласованности рассматриваемых переменных и их условная ранжировка по критерию степени тесноты связи каждой из них с остальными переменными.* Подобные задачи возникают, например, при исследовании степени согласованности мнений группы экспертов и при попытках условного упорядочения последних по их компетентности. В основе этого анализа лежит расчет коэффициента совокупной согласованности — *коэффициента конкордации* для различных комбинаций исследуемых переменных (см. п. 2.3).

Задача С: *построение единого группового упорядочения объектов на основе совокупности согласованных упорядочений «ядра»* (или нескольких групповых упорядочений — при наличии нескольких «ядер»). Решение этой задачи сводится к построению такого упорядочения, которое было бы, в определенном смысле, наиболее близким к каждому из упорядочений заданной совокупности — «ядра». Именно с такой задачей сталкивается, например, исследователь, желающий установить *неизвестное истинное упорядочение* заданной совокупности объектов по имеющемуся в его распоряжении набору экспертных ранжировок тех же объектов. Для построения единого (группового) варианта упорядочения $X^{(ед)}$ часто используют в качестве ранга $x_i^{(ед)}$ объекта O_i среднее арифметическое или медиану имеющихся базовых рангов $x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}$ этого объекта. Обоснование способа построения единого варианта упорядочения может быть получено, например, в рамках подхода, предложенного Дж. Кемени и Дж. Снеллом [63] (и распространенного затем Б. Г. Миркиным на случай *номинальных* признаков [92]), который опирается на введенную ими меру близости между ранжировками (определя-

ется ранжировка $X^{(д)}$, наименее удаленная , в смысле введенной меры близости, от всех ранжировок $X^{(1)}$, $X^{(2)}$, ..., $X^{(p)}$ базовой совокупности). Задача С может быть сформулирована и как задача наилучшего (в определенном смысле) восстановления ранжировки $X^{(0)}$, связанной с результирующей переменной $y \equiv x^{(0)}$, по ранжировкам $X^{(1)}$, $X^{(2)}$, ..., $X^{(p)}$, индуцируемым соответственно объясняющими переменными $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$. В такой формулировке ее называют также *задачей регрессии на порядковых (ординальных) переменных.*

2.1.4. Вероятностные пространства ранжировок, генерируемые порядковыми переменными [14, гл. 4, 5]. Вытекающая из определения порядковой случайной величины специфика заключается в первую очередь в том, что ее «возможные значения» определены в пространстве ранжировок, причем длина этих ранжировок (n) определяется числом статистически обследованных объектов (т. е. объемом выборки!). В то же время множество возможных значений *количественной* случайной переменной, а следовательно, и ее закон распределения вероятностей никак не зависят от объема обрабатываемой статистической выборки [14, гл. 5]. Для приведения «к общему знаменателю» этих двух схем можно воспользоваться одним из двух подходов·

а) формализованным описанием (с помощью той или иной математической модели) *самого механизма генерирования ранжировок*, основанным на допущении, что решение о предпочтении объекта O_i объекту O_j принимается на базе сравнения восстанавливаемых каким-либо способом *со случайной ошибкой* значений латентных (т. е. не поддающихся непосредственному измерению) числовых характеристик $v_i = v(O_i)$ и $v_j = v(O_j)$ «ценности» или «предпочтительности» этих объектов (см., например, о моделях Терстоуна—Мостеллера, Льюса и др. в кн.: Статистические методы анализа экспертных оценок. — М.: Наука, 1977); в этом случае параметр n (число сравниваемых объектов) сохраняет за собой роль объема выборки, а закон распределения вероятностей ранжировок рассматривается как распределение в выборочном пространстве, генерируемое вероятностным пространством случайной величины v;

б) определением в качестве i-го *случайного эксперимента* [14, п. 4.1.1] результата «наблюдения» $X^{(i)}$ ранжировки по i-му свойству ($i = 0, 1, 2, ..., p$); тогда число сравниваемых объектов n будет играть роль размерности нашего наблюдения, а объем выборки будет определяться числом рассматриваемых свойств (т. е. $p + 1$).

Остановимся на последнем подходе к построению и интерпретации вероятностных пространств ранжировок. В этом случае мы приходим к следующей модели вероятностного пространства ранжировок длины n, генерируемого порядковой переменной $x^{(k)}$[1]

Пространство элементарных исходов $\Omega = \{\omega_i\}_{i=\overline{1,M}}$ состоит из $M = n!$ всевозможных перестановок и не зависит от номера переменной k. Распределение вероятностей задается последовательностью $P^{(k)} = \{p_i^{(k)}\}_{i=\overline{1,M}}$, элементы которой, вообще говоря, зависят от номера «генерирующей» переменной k.

Поскольку множество элементарных исходов Ω дискретно (и конечно!), любое его подмножество измеримо и, следовательно, может быть интерпретировано как случайное событие.

Далее (см. § 2.2—2.3) будут предложены рекомендации по вычислению выборочных характеристик парной и множественной ранговой статистической связи. Однако исследование их важных статистических свойств (и в частности, конструирование на их основе статистических критериев и доверительных интервалов для неизвестных *теоретических* значений анализируемых характеристик) возможно лишь при некоторых дополнительных допущениях (гипотезах) относительно характера последовательностей $P^{(k)}$ и статистических связей между $x^{(0)}, x^{(1)}, \ldots, x^{(p)}$.

Наиболее исследованным является случай, когда постулируется справедливость следующей гипотезы H_0:

$$H_0: \begin{cases} \text{(а) случайные переменные } \{x^k\}_{k=\overline{0,p}} \text{ статистически} \\ \quad\ \text{независимы (см. [14, § 5.5]);} \\ \text{(в) все элементарные исходы равновероятны, т. е.} \end{cases} \tag{2.1}$$

$$p_1^{(k)} = p_2^{(k)} = \ldots = p_M^{(k)} = \frac{1}{n!},\quad k = 0, 1, \ldots, p.$$

Содержательно допущения гипотезы H_0 означают, что ранжирования заданного множества объектов по различным свойствам $x^{(0)}, x^{(1)}, \ldots, x^{(p)}$ никак друг с другом не связаны (допущение (а)) и что ни одно из этих свойств не определяет никаких предпочтений в задаче сравнения «качества» анализируе-

[1]Для упрощения обозначений здесь рассматривается лишь случай *строгих* упорядочений, т. е. ситуации, когда принципиально невозможна неразличимость рангов (объединение рангов). Общий случай имеет лишь технические отличия, связанные с увеличением общего числа M элементарных исходов.

мых объектов, так как в результате случайного эксперимента с *одинаковой вероятностью* может появиться любое из $n!$ возможных упорядочений (допущение (в)). К сожалению, статистический анализ, проведенный в рамках допущений (2.1), дает возможность лишь принять или отклонить гипотезу H_0. А поскольку на практике выборочные ранговые корреляционные характеристики оказываются, как правило, весьма высокими по абсолютной величине (что свидетельствует о том, что мы находимся вне условий нулевой гипотезы), то их распределение в реальной ситуации оказывается неизвестным и на их основе не удается делать дальнейшие выводы (аналогичные, например, тем, которые следуют из п. 1.1.3, 1.1.5, 1.2.3, 1.3.3 относительно парных, частных и множественных корреляционных связей между количественными переменными).

Более интересными в прикладном плане нам представляются условия, постулируемые в рамках гипотезы H_1:

H_1:
- (а) случайные переменные $\{x^{(k)}\}_{k=\overline{0,p}}$ статистически независимы;
- (в′) случайные переменные $\{x^{(k)}\}_{k=\overline{0,p}}$ одинаково распределены, т. е. $P^{(0)} = P^{(1)} = \ldots = P^{(p)} = P$;
- (с) распределение вероятностей P обладает свойством *монотонности* относительно некоторого истинного упорядочения ω_{i_0}. (2.2)

Под свойством *монотонности* понимается выполнение следующего условия: если введенное некоторым образом «расстояние» $d(\omega_i, \omega_{i_0})$ между любым упорядочением ω_i и некоторым «истинным» упорядочением ω_{i_0} не превосходит $d(\omega_{i'}, \omega_{i_0})$, то $p_i \geqslant p_{i'}$; другими словами, чем «ближе» ранжировка к истинной, тем с большей вероятностью мы ее получим в результате случайного эксперимента над переменной $x^{(k)}$ $(k = 0, 1, \ldots, p)$, и, следовательно, истинная ранжировка ω_{i_0} является наиболее вероятным исходом случайного эксперимента.

Некоторые результаты, связанные со статистическим анализом ранжировок в рамках условий (2.2), можно найти в [105, 131].

2.2. Анализ и измерение парных ранговых статистических связей

2.2.1. Ранговый коэффициент корреляции Спирмэна. Для измерения степени тесноты связи между ранжировками $X_{\cdot}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)})'$ и $X_{\cdot}^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)})'$

К. Спирмэн еще в 1904 г. предложил показатель

$$\widehat{\tau}_{kj}^{(S)} = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (x_i^{(k)} - x_i^{(j)})^2, \tag{2.3}$$

названный впоследствии *ранговым коэффициентом корреляции Спирмэна.* Прямым подсчетом нетрудно убедиться, что для *совпадающих* ранжировок (т. е. при $x_i^{(k)} = x_i^{(j)}$ для всех $i = 1, 2, \ldots, n$) $\widehat{\tau}_{kj}^{(s)} = 1$, а *для противоположных* (т. е. при $x^{(k)} = n - x_i^{(j)} + 1$, $i = 1, 2, \ldots, n$) $\widehat{\tau}_{kj}^{(s)} = -1$. Можно показать (см., например, [67]), что во всех остальных случаях $|\widehat{\tau}_{kj}^{(s)}| < 1$.

Формула (2.3) пригодна лишь в случае отсутствия объединенных рангов в обеих исследуемых ранжировках. Для ее распространения на общий случай определим для каждой (k-й) ранжировки $X_{\cdot}^{(k)}$ ($k = 0, 1, \ldots, p$) величину

$$T^{(k)} = \frac{1}{12} \sum_{t=1}^{m^{(k)}} [(n_t^{(k)})^3 - n_t^{(k)}], \tag{2.4}$$

где $m^{(k)}$ — число групп неразличимых рангов у переменной $x^{(k)}$, а $n_t^{(k)}$ — число элементов (рангов), входящих в t-ю группу неразличимых рангов (в частном случае отсутствия объединенных рангов имеем $m^{(k)} = n$, $n_1^{(k)} = n_2^{(k)} = \ldots = n_n^{(k)} = 1$ и соответственно $T^{(k)} = 0$; кроме того, группы неразличимых рангов, состоящие из единственного элемента, по существу, не участвуют в расчете величины $T^{(k)}$).

Тогда ранговый коэффициент корреляции Спирмэна между ранжировками $X_{\cdot}^{(k)}$ и $X_{\cdot}^{(j)}$ следует вычислять по формуле

$$\widehat{\tau}_{kj}^{(S)} = \frac{\frac{1}{6}(n^3 - n) - \sum_{i=1}^{n} (x_i^{(k)} - x_i^{(j)})^2 - T^{(k)} - T^{(j)}}{\sqrt{\left[\frac{1}{6}(n^3 - n) - 2T^{(k)}\right]\left[\frac{1}{6}(n^3 - n) - 2T^{(j)}\right]}}. \tag{2.5}$$

Если $T^{(k)}$ и $T^{(j)}$ являются небольшими относительно $\frac{1}{6}(n^3 - n)$ величинами, то можно воспользоваться приближенным соотношением (а при $T^{(k)} = T^{(j)}$ оно точное)

$$\widehat{\tau}_{kj}^{(S)} = 1 - \frac{\sum_{i=1}^{n} \left(x_i^{(k)} - x_i^{(j)}\right)^2}{\frac{1}{6}(n^3 - n) - (T^{(k)} + T^{(j)})} \qquad (2.5')$$

Правда, при этом же условии (относительная малость $T^{(k)} + T^{(j)}$ по сравнению с $\frac{1}{6}(n^3 - n)$) и приближенная формула (2.3) дает хорошую точность.

Пример 2.1. Два эксперта проранжировали 10 предложенных им проектов реорганизации научно-производственного объединения (НПО) с точки зрения их эффективности (при заданных ресурсных ограничениях). Занумеровав проекты в порядке ранжировки 1-го эксперта, получаем в качестве исходных данных: $X^{(1)'} = (1;\ 2;\ 3;\ 4;\ 5;\ 6;\ 7;\ 8;\ 9;\ 10)$; $X^{(2)'} = (2;\ 3;\ 1;\ 4;\ 6;\ 5;\ 9;\ 7;\ 8;\ 10)$.

Вычисления по формуле (2.3) дают:

$$\tau_{12}^{(S)} = 1 - \frac{6}{1000 - 10} \cdot (1 + 1 + 2^2 + 0 + 1^2 + 1^2 + 2^2 + 1 + 1 + 0) = 1 - \frac{6}{990} \cdot 14 = 0{,}915,$$

что свидетельствует о существенной положительной ранговой связи между исследуемыми переменными.

Пример 2.2. Десять однородных предприятий подотрасли были проранжированы вначале по степени прогрессивности их оргструктур (признак $x^{(1)}$), а затем — по эффективности их функционирования в отчетном году (признак $x^{(2)}$). В результате были получены следующие две ранжировки: $X^{(1)'} = (1;\ 2{,}5;\ 2{,}5;\ 4{,}5;\ 4{,}5;\ 6{,}5;\ 6{,}5;\ 8;\ 9{,}5;\ 9{,}5)$; $X^{(2)'} = (1;\ 2;\ 4{,}5;\ 4{,}5;\ 4{,}5;\ 4{,}5;\ 8;\ 8;\ 8;\ 10)$.

В первой ранжировке имеем четыре группы неразличимых рангов, число элементов в которых больше единицы, а во второй ранжировке — две такие группы. В соответствии с формулой (2.4) получаем:

$$T^{(1)} = \frac{1}{12}[(2^3 - 1) + (2^3 - 1) + (2^3 - 1) + (2^3 - 1)] = \frac{28}{12} = 2{,}33;$$

$$T^{(2)} = \frac{1}{12}[(4^3 - 1) + (3^3 - 1)] = 7{,}42.$$

Точная формула (2.5) дает $\widehat{\tau}_{12}^{(s)} = 0{,}917$. Вычисление этого же коэффициента корреляции по *приближенным* фор-

мулам (2.3) и (2.5′) дает соответственно значения 0,921 и 0,917. Все эти результаты оказываются совпадающими при округлении до второго десятичного знака.

2.2.2. Ранговый коэффициент корреляции Кендалла. Другой широко используемой характеристикой тесноты статистической связи между двумя упорядочениями является ранговый коэффициент корреляции Кендалла, определяемый соотношением [67]

$$\widehat{\tau}_{kj}^{(K)} = 1 - \frac{4\nu (X_{\cdot}^{(k)},\ X_{\cdot}^{(j)})}{n(n-1)}, \tag{2.6}$$

где $\nu\ (X_{\cdot}^{(k)},\ X_{\cdot}^{(j)})$ — минимальное число обменов соседних элементов последовательности $X_{\cdot}^{(j)}$, необходимое для приведения ее к упорядочению $X_{\cdot}^{(k)}$. Очевидно, величина $\nu\ (X_{\cdot}^{(k)},\ X_{\cdot}^{(j)})$ симметрична относительно своих аргументов, так что с равным правом можно говорить о минимальном числе «соседских обменов» элементов последовательности $X_{\cdot}^{(k)}$, необходимом для приведения ее к виду $X_{\cdot}^{(j)}$.

Из (2.6) сразу следует, что при совпадающих ранжировках $X_{\cdot}^{(k)}$ и $X_{\cdot}^{(j)}$ $\widehat{\tau}_{kj}^{(K)} = 1$ (так как $\nu\ (X_{\cdot}^{(k)},\ X_{\cdot}^{(j)} = 0)$, а при противоположных (т. е. при $x_i^{(k)} = n - x_i^{(j)} + 1$, $i = 1, 2, \ldots, n$, так что $\nu\ (X_{\cdot}^{(k)},\ X_{\cdot}^{(j)}) = \frac{1}{2}\ n\ (n-1)$) $\widehat{\tau}_{jk}^{(K)} = -1$. Нетрудно показать (см., например, [67]), что во всех остальных случаях $|\tau_{kj}^{(K)}| < 1$.

Вычисление $\widehat{\tau}_{kj}^{(K)}$ связано с необходимостью подсчета величины $\nu\ (X_{\cdot}^{(k)},\ X_{\cdot}^{(j)})$ и, следовательно, является более *трудоемким*, чем вычисление $\tau_{kj}^{(S)}$. Однако, во-первых, коэффициент Кендалла обладает некоторыми преимуществами по сравнению с коэффициентом Спирмэна, главные из них: а) относительно бо́льшая продвинутость в исследовании его статистических свойств и, в частности, его выборочного распределения (см. ниже, п. 2.2.4); б) возможность его использования и в частной («очищенной») корреляции рангов [67, гл. 8]); в) бо́льшие удобства его пересчета при добавлении к n статистически обследованным объектам новых, т. е. при удлинении анализируемых ранжировок: для вычисления нового значения рангового коэффициента корреляции приходится переранжировать значительную часть объектов, что в случае $\tau_{kj}^{(S)}$ означает необходимость пересчета разностей $x_i^{(k)} - x_i^{(j)}$; при вычислении же $\tau_{ij}^{(K)}$ значения рангов не играют никакой роли, важно лишь число необходимых «сосед-

ских обменов», которое при добавлении новых объектов подсчитывается рекуррентным способом (к старому значению $\nu(X^{(k)}_{\cdot}, X^{(j)}_{\cdot})$ может быть лишь дополнен некоторый «добавок»).

Во-вторых, можно воспользоваться рекомендациями, упрощающими подсчет числа $\nu(X^{(k)}_{\cdot}, X^{(j)}_{\cdot})$ как при ручном, так и при машинном счете.

Так, при ручном счете полезным оказывается известный факт тождественного совпадения величин $\nu(X^{(k)}_{\cdot}, X^{(j)}_{\cdot})$ и $I(X^{(k)}_{\cdot}, X^{(j)}_{\cdot})$, где *число инверсий* $I(X^{(k)}_{\cdot}, X^{(j)}_{\cdot})$ — это просто число расположенных в неодинаковом порядке пар элементов последовательностей $X^{(k)}_{\cdot}$ и $X^{(j)}_{\cdot}$, являющееся естественной мерой нарушения порядка объектов в одной последовательности относительно другой. Для удобства подсчета $I(X^{(k)}_{\cdot}, X^{(j)}_{\cdot})$ перенумеруем объекты в порядке, определяемом рангами последовательности $X^{(k)}_{\cdot}$. Тогда анализируемые ранжировки $X^{(k)}_{\cdot}$, $X^{(j)}_{\cdot}$ соответствующим образом видоизменяются, т. е. преобразуются к виду соответственно $\widetilde{X}^{(k)}_{\cdot}$, $\widetilde{X}^{(j)}_{\cdot}$, где $\widetilde{X}^{(k)\prime}_{\cdot} = (1, 2, \ldots, n)$; $\widetilde{X}^{(j)\prime}_{\cdot} = (\widetilde{x}^{(j)}_1, \widetilde{x}^{(j)}_2, \ldots, \widetilde{x}^{(j)}_n)$, а число инверсий $I(X^{(k)}_{\cdot}, X^{(j)}_{\cdot}) \equiv I(\widetilde{X}^{(k)}_{\cdot}, \widetilde{X}^{(j)}_{\cdot})$, а следовательно, и величина $\nu(X^{(k)}_{\cdot}, X^{(j)}_{\cdot})$ определятся по формуле

$$\nu(X^{(k)}_{\cdot}, X^{(j)}_{\cdot}) = I(X^{(k)}_{\cdot}, X^{(j)}_{\cdot}) = \sum_{q=1}^{n-1} \sum_{l=q+1}^{n} \nu^{(j,k)}_{ql}, \qquad (2.7)$$

где

$$\nu^{(j,k)}_{ql} = \begin{cases} 1, \text{ если } \widetilde{x}^{(j)}_q > \widetilde{x}^{(j)}_l \text{ (т. е. нарушен порядок последовательности } \widetilde{X}^{(k)}_{\cdot}) \\ 0 \text{ — в противоположном случае.} \end{cases}$$

Легко подсчитать, что число инверсий $I(X^{(k)}_{\cdot}, X^{(j)}_{\cdot})$ может меняться от 0 (что соответствует случаю совпадающих ранжировок) до $\frac{1}{2}n(n-1)$ (что соответствует случаю противоположных ранжировок).

Формулы (2.6)—(2.7) пригодны для подсчета $\widehat{\tau}^{(K)}_{kj}$ лишь в случае отсутствия объединенных рангов в обеих исследуемых ранжировках. Соответствующее «подправленное» значение $\widehat{\tau*}^{(K)}_{kj}$ при наличии объединенных рангов в анализируемых упорядочениях будет определяться соотношением

$$\widehat{\tau}^{*(K)}_{kj} = \frac{\widehat{\tau}^{(K)}_{kj} - \frac{2(U^{(1)}+U^{(2)})}{n(n-1)}}{\sqrt{\left(1-\frac{2U^{(k)}}{n(n-1)}\right)\left(1-\frac{2U^{(j)}}{n(n-1)}\right)}} \qquad (2.6')$$

в котором коэффициент $\widehat{\tau}^{(K)}_{kj}$ вычисляется по формуле (2.6)—(2.7), а «поправочные» величины $U^{(l)}$ определяются соотношением

$$U^{(l)} = \frac{1}{2}\sum_{t=1}^{m^{(l)}} n_t^{(l)}\left(n_t^{(l)}-1\right),\ l=k,\ j \qquad (2.8)$$

(смысл величин $m^{(l)}$ и $n_t^{(l)}$ определен в п. 2.2.1, см. (2.4)).

Для пояснения работоспособности формул (2.6)—(2.8) вернемся к примерам 2.1, 2.2.

Анализ степени согласованности ранжировок двумя экспертами десяти проектов реорганизации НПО (пример 2.1), осуществленный с использованием формул (2.6), (2.7), дает:

$$\nu_{12}=0;\ \nu_{13}=1,\ \nu_{14}=\nu_{15}=\nu_{16}=\nu_{17}=\nu_{18}=\nu_{19}=\nu_{1.10}=0;$$
$$\nu_{23}=1;\ \nu_{24}=\nu_{25}=\nu_{26}=\nu_{27}=\nu_{28}=\nu_{29}=\nu_{2.10}=0;$$
$$\nu_{34}=1;\ \nu_{35}=\nu_{36}=\nu_{37}=\nu_{38}=\nu_{39}=\nu_{3.10}=0;$$
$$\nu_{45}=\nu_{46}=\nu_{47}=\nu_{48}=\nu_{49}=\nu_{4.10}=0;$$
$$\nu_{56}=1;\ \nu_{57}=\nu_{58}=\nu_{59}=\nu_{5.10}=0;$$
$$\nu_{67}=\nu_{68}=\nu_{69}=\nu_{6.10}=0;$$
$$\nu_{78}=1;\quad \nu_{79}=1;\quad \nu_{7.10}=0;$$
$$\nu_{89}=\nu_{8.10}=0;$$
$$\nu_{9.10}=0.$$

Таким образом, $\nu(X^{(1)}_{\cdot}, X^{(2)}_{\cdot}) = 1+1+1+0+1+0+ +2+0+0=6$.

Соответственно

$$\widehat{\tau}^{(K)}_{12} = 1-\frac{4\cdot 6}{10\cdot 9} = 1-0{,}267 = 0{,}733$$

(напомним, что коэффициент Спирмэна в этом примере был равным 0,915).

При вычислении рангового коэффициента корреляции Кендалла в примере 2.2 следует воспользоваться формулой (2.6′), так как исследуемые ранжировки содержат объединенные ранги. Используя результаты расчета величин $m^{(1)} = 4$, $m^{(2)} = 2$, $n_1^{(1)} = n_2^{(1)} = n_3^{(1)} = n_4^{(1)} = 2$, $n_1^{(2)} = 4$, $n_2^{(2)} = 3$ (см. п. 2.2.1), получаем (в соответствии с (2.8)):

$$U^{(1)} = \frac{1}{2}(2+2+2+2) = 4; \quad U^{(2)} = \frac{1}{2}(4\cdot 3 + 3\cdot 2) = 9.$$

Обращаясь теперь к формуле (2.6′), имеем:

$$\widehat{\tau}^{*(K)}_{12} = \frac{1 - \frac{26}{90}}{\sqrt{\left(1 - \frac{8}{90}\right)\left(1 - \frac{18}{90}\right)}} = 0{,}833$$

(напомним, что соответствующий коэффициент Спирмэна был равен 0,917).

2.2.3. Обобщенная формула для парного коэффициента корреляции и связь между коэффициентами Спирмэна и Кендалла. Для удобства стандартной реализации системы алгоритмов корреляционного анализа на ЭВМ полезно ввести некоторый *обобщенный* прием вычисления парных корреляционных характеристик, определенный для любой двумерной системы n наблюдений

$$\begin{pmatrix} X^{(k)\prime}_{\cdot} \\ X^{(j)\prime}_{\cdot} \end{pmatrix} = \begin{pmatrix} x_1^{(k)}, & x_2^{(k)}, & \ldots, & x_n^{(k)} \\ x_1^{(j)}, & x_2^{(j)}, & \ldots, & x_n^{(j)} \end{pmatrix}. \tag{2.9}$$

С этой целью определим некоторое правило, в соответствии с которым каждой паре $(x_{i_1}^{(l)}, x_{i_2}^{(l)})$ компонент вектора $X^{(l)}_{\cdot}$ $(l = k, j)$ ставится в соответствие число («метка») $a_{i_1 i_2}^{(l)}$, причем это правило должно обладать свойством *отрицательной симметричности* (т. е. $a_{i_1 i_2}^{(l)} = -a_{i_2 i_1}^{(l)}$) и центрированности (т. е. $a_{ii}^{(l)} = 0$ при всех $l = k, j$ и всех $i = 1, 2, \ldots, n$). Тогда *обобщенный коэффициент корреляции* $r^{(об)}$ переменных $x^{(k)}$ и $x^{(j)}$ определяется формулой

$$\widehat{r}_{kj}^{(об)} = \frac{\sum\limits_{i_1=1}^{n}\sum\limits_{i_2=1}^{n} a_{i_1 i_2}^{(k)} a_{i_1 i_2}^{(j)}}{\sqrt{\sum\limits_{i_1=1}^{n}\sum\limits_{i_2=1}^{n}(a_{i_1 i_2}^{(k)})^2 \cdot \sum\limits_{i_1=1}^{n}\sum\limits_{i_2=1}^{n}(a_{i_1 i_2}^{(j)})^2}}. \tag{2.10}$$

Легко видеть (см., например, [67]), что практически все введенные нами характеристики парной корреляционной свя-

зи могут быть получены как частные случаи формулы (2.10) при соответствующем выборе правила приписывания числовых «меток» $a_{i_1 i_2}$. Действительно:

а) положив $a_{i_1 i_2}^{(l)} = x_{i_1}^{(l)} - x_{i_2}^{(l)}$, $l = k, j$, получаем формулу для *обычного парного коэффициента корреляции* $\widehat{r}_{kj}$, если $x_i^{(l)}$ — значение l-й количественной переменной в i-м наблюдении (см. п. 1.1.2, формулу (1.8')), и формулу для рангового коэффициента корреляции Спирмэна $\widehat{\tau}_{kj}^{(S)}$, если $x_i^{(l)}$ — ранг i-го объекта в ряду, упорядоченном по порядковой переменной $x^{(l)}$ (см. формулу (2.3));

б) положив

$$a_{i_1 i_2}^{(l)} = \begin{cases} +1, \text{ если } x_{i_1}^{(l)} < x_{i_2}^{(l)}; \\ 0, \text{ если } x_{i_1}^{(l)} = x_{i_2}^{(l)}; \\ -1, \text{ если } x_{i_1}^{(l)} > x_{i_2}^{(l)}, \end{cases}$$

получаем формулы (2.6) и (2.6') для рангового коэффициента корреляции Кендалла $\widehat{\tau}_{kj}^{(K)}$, если под $x_i^{(l)}$ понимать ранг i-го объекта в l-м упорядочении.

Заметим, что значения ранговых корреляционных характеристик $\widehat{\tau}_{kj}^{(K)}$ и $\widehat{\tau}_{kj}^{(S)}$ довольно тесно связаны одно с другим. Это следовало ожидать, так как *обе характеристики являются линейными функциями от числа инверсий*, имеющихся в сравнении последовательностей $X_{\cdot}^{(k)}$ и $X_{\cdot}^{(j)}$: различие этих функций состоит в том, что при подсчете коэффициента Спирмэна инверсиям более отдаленных (по величине) друг от друга элементов приписываются большие веса (см., например, [67, п. 1.17 и 2.12]). Между масштабами шкал, в которых измеряют корреляцию коэффициенты $\widehat{\tau}^{(S)}$ и $\widehat{\tau}^{(K)}$ нет простого соотношения. Однако уже при умеренно больших значениях n ($n \geqslant 10$) и при условии, что абсолютные величины значений этих коэффициентов не слишком близки к единице, их связывает следующее простое приближенное соотношение

$$\widehat{\tau}^{(S)} \approx 1{,}5\, \widehat{\tau}^{(K)}.$$

2.2.4. Статистические свойства выборочных характеристик парной ранговой связи. До сих пор речь шла о *выборочных* характеристиках ранговой связи. Попробуем ответить на вопрос: как точно эти выборочные характеристики (определенные, в частности, формулами (2.3)—(2.8)) оценивают соответствующие истинные (*теоретические*) значения?

Для этого в первую очередь следует пояснить, что в данном случае понимается под теоретическими характеристиками.

Представим себе сначала *конечную* генеральную совокупность, состоящую из N объектов $O_1, O_2, \ldots, O_N$, каждый из которых снабжен двумя порядковыми номерами: $O_i \leftrightarrow (x_i^{(k)}, x_i^{(j)})$, $i = 1, 2, \ldots, N$, где $x_i^{(l)}$ означает место объекта O_i в общем ряду всех N объектов, упорядоченном по степени выраженности свойства $x^{(l)}$ $(l = k, j)$. Будем полагать, что статистически обследованное множество объектов $O_{i_1}, O_{i_2}, \ldots, O_{i_n}$ образуется как случайная выборка объема n, взятая из совокупности $O_1, O_2, \ldots, O_N$ $(n \ll N)$.

Определим *теоретические* (истинные) значения коэффициентов $\tau_{kj}^{(S)}$, $\tau_{kj}^{(K)}$ и $r_{kj}^{(об)}$ соответственно теми же соотношениями (2.3) (или (2.5)), (2.6) (или 2.6')) и (2.10), что и выборочные с заменой объема выборки n объемом генеральной совокупности N. При работе с выборкой производится *естественная перенумерация* объектов и их рангов, не меняющая их упорядоченности в генеральной совокупности ни по одной из переменных.

В дальнейшем нас будет интересовать, как сильно могут отличаться выборочные значения $\widehat{\tau}^{(S)}$ и $\widehat{\tau}^{(K)}$ от соответствующих теоретических, в том числе в так называемых *асимптотических ситуациях*, т. е. при $N \to \infty$ и $n(N) \to \infty$.

Проверка статистически значимого отличия от нуля ранговых корреляционных характеристик (т. е. проверка гипотезы H_0, см. соотношения (2.1)) осуществляется при «не слишком малых» n (т. е. при $n > 10$) при заданном уровне значимости критерия α с помощью проверки неравенств

$$\left|\widehat{\tau}^{(S)}\right| > t_{\frac{\alpha}{2}}(n-2) \cdot \sqrt{\frac{1-\left(\widehat{\tau}^{(S)}\right)^2}{n-2}}; \qquad (2.11)$$

$$\left|\widehat{\tau}^{(K)}\right| > u_{\frac{\alpha}{2}} \cdot \sqrt{\frac{2(2n+5)}{9n(n-1)}}, \qquad (2.12)$$

в которых $t_q(\nu)$ и u_q, как и прежде, $100q\%$-ные точки соответственно $t(\nu)$- и нормального распределения (см. табл. П.6 и П.3). Выполнение неравенств (2.11) и (2.12) сигнализирует о необходимости отвергнуть гипотезу H_0, т. е. о наличии статистически значимой ранговой корреляционной связи. В случае небольших объемов выборок $(4 \leqslant n \leqslant 10)$ статистическая проверка гипотезы об отсутствии ранговой корреляционной связи производится с помощью табл. П.9 и П.10.

Таблица П.9 позволяет при малых n ($n = 4, 5, \ldots, 10$) построить то пороговое значение $\tau_{max}^{(S)}$, при превышении которого (по абсолютной величине) коэффициентом Спирмэна $\widehat{\tau}^{(S)}$ следует признать наличие статистически значимой связи между анализируемыми переменными. Задавшись уровнем значимости критерия α и числом сравниваемых объектов n, определяем из таблицы величину $S_C = S_C(n, Q)$, соответствующую нашему n и значению $Q = \alpha/2$ (или *приблизительно* равному $\alpha/2$). Тогда

$$\tau_{max}^{(S)} = \frac{2S_C(n, Q)}{K_n} - 1, \tag{2.13}$$

где $K_n = \frac{1}{3}(n^3 - n)$ (значения этой вспомогательной константы приведены в последней строке таблицы).

Так, в примере 2.1 для уровня значимости $\alpha = 0{,}06$ имеем: $n = 10$; $Q = 0{,}03$; $S_C = S_C(10;\ 0{,}3) = 268$; $K_{10} = 330$, так что в соответствии с (2.13)

$$\tau_{max}^{(S)} = \frac{2 \cdot 268}{330} - 1 = 0{,}624.$$

Поскольку выборочное значение рангового коэффициента корреляции Спирмэна $\tau^{(S)}$ в этом примере значительно превосходит пороговое значение ($\widehat{\tau}^{(S)} = 0{,}915 > 0{,}624$), то гипотеза об отсутствии корреляционной связи отвергается.

И наконец, в табл. П.10 приведены значения вспомогательных величин S_K, позволяющих вычислить (при малых $n = 4, 5, \ldots, 10$) то пороговое значение $\tau_{max}^{(K)}$, при превышении которого (по абсолютной величине) коэффициентом Кендалла следует признать наличие статистически значимой связи между анализируемыми переменными. Для этого поступают следующим образом: задавшись объемом выборки n и уровнем значимости критерия α, находят в столбце, соответствующем данному n, величину, равную (или приблизительно равную) $\alpha/2$; затем находят значение $S_K = S_K(n, \alpha)$ в левом столбце *той же самой строки* и вычисляют $\tau_{max}^{(K)}$ по формуле

$$\tau_{max}^{(K)} = \frac{2S_K(n, \alpha)}{n \cdot (n-1)}. \tag{2.14}$$

Если окажется, что $\widehat{\tau}^{(K)} > \tau_{max}^{(K)}$, то гипотеза об отсутствии ранговой корреляционной связи отвергается (связь статистически значима).

Так, в примере 2.1 при уровне значимости $\alpha = 0{,}06$ имеем: $n = 10$; $0{,}23 < \frac{\alpha}{2} < 0{,}36$; следовательно, $S_K = 22$ (оно лежит между 21 и 23), так что

$$\tau_{\max}^{(K)} = \frac{2 \cdot 22}{10 \cdot 9} = \frac{44}{90} = 0{,}489.$$

Поскольку $\tau^{(K)} = 0{,}733 > 0{,}489$, делается вывод о наличии статистически значимой корреляционной связи между исследуемыми переменными в данном примере.

Построение доверительных интервалов для неизвестных истинных значений ранговых коэффициентов корреляции возможно лишь приближенно и только при измерении ранговой корреляции с помощью коэффициента Кендалла. При этом используют (при $n > 10$ и значениях $\tau^{(K)}$, не слишком близких по абсолютной величине к единице) приближенный факт нормальности распределения величины $\widehat{\tau}^{(K)}$ со средним значением $\mathbf{E}\widehat{\tau}^{(K)} \approx \tau^{(K)}$ и с дисперсией $\widehat{\mathbf{D}\tau}^{(K)}$, не превышающей величины $\frac{2}{n}[1 - (\tau^{(K)})^2]$. Можно утверждать, что с доверительной вероятностью, не меньшей заданного уровня P, истинное значение коэффициента Кендалла $\tau^{(K)}$ заключено в пределах

$$\widehat{\tau}^{(K)} - u_{\frac{1+P}{2}} \cdot \sqrt{\frac{2}{n}[1 - (\widehat{\tau}^{(K)})^2]} < \tau^{(K)} < \widehat{\tau}^{(K)} + u_{\frac{1+P}{2}} \cdot \sqrt{\frac{2}{n}[1 - (\widehat{\tau}^{(K)})^2]}, \qquad (2.15)$$

где u_q — q-квантиль стандартного нормального распределения (см. табл. П.3).

2.3. Анализ множественных ранговых связей

2.3.1. Коэффициент конкордации (согласованности) как измеритель статистической связи между несколькими порядковыми переменными. До сих пор мы рассматривали корреляцию между двумя порядковыми переменными. Однако при решении основных задач A—C статистического анализа ранговых связей (см. п. 2.1.3) возникает необходимость уметь измерить статистическую связь между *несколькими* (более чем двумя) переменными. С этой целью Кендаллом [67] был предложен показатель $\widehat{W}(m)$, названный *коэффициентом кон-*

кордации (или *согласованности*), вычисляемый по формуле[1]

$$\widehat{W}(m) = \frac{12}{m^2(n^3-n)} \cdot \sum_{i=1}^{n} \left(\sum_{j=1}^{m} x_i^{(k_j)} - \frac{m(n+1)}{2} \right)^2, \qquad (2.16)$$

где m — число анализируемых порядковых переменных (сравниваемых упорядочений); n — число статистически обследованных объектов или длина ранжировки (объем выборки); k_1, k_2, ..., k_m — номера отобранных для анализа порядковых переменных (из исходной совокупности $x^{(0)}$, $x^{(1)}$, $x^{(2)}$, ..., $x^{(p)}$, так что, очевидно, $m \leqslant p + 1$).

Нетрудно устанавливаются следующие свойства коэффициента конкордации (см., например, [67, гл. 6]):

а) $0 \leqslant \widehat{W} \leqslant 1$;

б) $\widehat{W} = 1$ тогда и только тогда, когда все m анализируемых упорядочений совпадают;

в) если $m \geqslant 3$ и анализируемые ранжировки генерируются подобно случайному независимому m-кратному извлечению из множества всех $n!$ возможных упорядочений n объектов (условия гипотезы H_0, см. п. 2.1.4), то связи между ними нет и $W = 0$;

г) пусть $\bar{\tau}^{(S)}(m)$ — среднее значение коэффициента Спирмэна, подсчитанное по значениям $m(m-1)/2$ коэффициентов $\widehat{\tau}^{(S)}_{k_i k_j}$ ($i, j = 1, 2, ..., m$; $i \neq j$), характеризующих ранговую связь между всеми возможными парами переменных $(x^{(k_i)}, x^{(k_j)})$ из анализируемого набора $(x^{(k_1)}, x^{(k_2)}, ..., x^{(k_m)})$; тогда

$$\bar{\tau}^{(S)}(m) = \frac{m\widehat{W}(m) - 1}{m-1}; \qquad (2.17)$$

в частности, из (2.17) следует для случая $m = 2$, что

$$\widehat{W}(2) = \frac{1}{2}\left(\widehat{\tau}^{(S)}_{k_1 k_2} + 1\right), \qquad (2.17')$$

т. е. коэффициент конкордации, исчисленный для *двух* переменных, пропорционален введенному ранее парному ранговому коэффициенту корреляции Спирмэна.

---

[1]Мы приводим здесь формулу для подсчета *выборочного* значения $\widehat{W}$ коэффициента конкордации W. Интерпретация и вычисление *теоретического* значения W непосредственно следуют из рассуждений, приведенных в п. 2.2.4 в связи с анализом статистических свойств выборочных парных ранговых коэффициентов корреляции.

То, что шкала измерения $W(m)$ не включает в себя отрицательных значений, объясняется следующим обстоятельством. В отличие от случая парных связей при анализе m ($m \geqslant 3$) порядковых переменных противоположные понятия согласованности и несогласованности утрачивают прежнюю симметричность (относительно нуля); упорядочения, произведенные в соответствии с переменными $x^{(k_1)}$, $x^{(k_2)}$, ..., $x^{(k_m)}$, могут полностью совпадать, *но не могут полностью не совпадать* в том смысле, который мы вкладывали в это понятие при $m = 2$.

Формула (2.16) получена (и справедлива) в предположении отсутствия объединенных рангов в каждом из анализируемых упорядочений. Если же таковые имеются, то формула должна быть модифицирована:

$$\widehat{W}(m) = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{m} x_i^{(k_j)} - \frac{m(n+1)}{2}\right)^2}{\frac{1}{12} m^2 (n^3 - n) - m \sum_{j=1}^{m} T^{(k_j)}}, \qquad (2.16')$$

где поправочный коэффициент $T^{(k_j)}$ (соответствующий переменной $x^{(k_j)}$) подсчитывается по формуле (2.4).

2.3.2. Проверка статистической значимости выборочного значения коэффициента конкордации. Как ведут себя выборочные значения $\widehat{W}(m)$ коэффициента конкордации при повторении выборок заданного объема n (из одной и той же генеральной совокупности) при отсутствии какой-либо связи между анализируемыми m переменными? Другими словами, нас интересует ответ на следующий вопрос. Предположим, что каждому объекту конечной *генеральной совокупности* (состоящей из N элементов) приписан какой-то определенный ранг по каждой из m рассматриваемых переменных. Так, например, если $m = 3$ и объекту O_i приписана тройка ($x_i^{(1)} = N$; $x_i^{(2)} = 1$; $x_i^{(3)} = 2$), то это означает, что по переменной $x^{(1)}$ он стоит на последнем (N-м) месте в упорядоченном ряду *всех* объектов генеральной совокупности, по переменной $x^{(2)}$ — на первом и по переменной $x^{(3)}$ — на втором. Тогда по исходным данным $\{(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(m)})\}_{i=\overline{1,N}}$ с помощью формулы (2.16) может быть вычислен теоретический (генеральный) коэффициент конкордации $W(m)$, характеризующий степень тесноты ранговой связи между переменными $x^{(1)}$, $x^{(2)}$, ..., $x^{(m)}$. Однако исследователю известны значения ($x_i^{(1)}$, $x_i^{(2)}$,..., $x_i^{(m)}$) *лишь для части объектов* генеральной совокупности, а именно для слу-

чайной выборки объектов объема n ($n < N$). После естественной перенумерации рангов, сохраняющей правило упорядочения объектов, но переводящей масштаб измерения рангов в шкалу (1, 2, ..., n) (для этого минимальный из оказавшихся в выборке рангов по каждой переменной объявляется рангом, равным 1, следующий по величине — рангом, равным 2, и т. д.), может быть вычислен (по той же формуле (2.16)) *выборочный* коэффициент конкордации $\widehat{W}(m)$. Извлекая другую выборку объема n из той же самой генеральной совокупности, мы получим, вообще говоря, другое значение выборочного коэффициента $\widehat{W}(m)$ и т. д.

Спрашивается, как сильно могут отклоняться от нуля выборочные значения коэффициента конкордации $\widehat{W}(m)$ в ситуации, когда значение теоретического коэффициента конкордации $W(m)$ свидетельствует о полном отсутствии ранговой связи между анализируемыми переменными $x^{(1)}$, $x^{(2)}, \ldots, x^{(m)}$? Для малых значений m и n ($2 \leqslant m \leqslant 20$, $3 \leqslant n \leqslant 7$) ответ на этот вопрос может быть получен с помощью табл. П.11а. Обозначенная в ней величина S есть не что иное, как

$$S = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} x_i^{(k_j)} - \frac{m(n+1)}{2} \right)^2. \qquad (2.18)$$

«Входами» в табл. П.11а является тройка чисел (m, n, S). «Выходом» — вероятность того, что величина S может быть такой, какой она является в нашей выборке, или большей в условиях отсутствия связи переменных в генеральной совокупности. Если окажется, что эта вероятность меньше принятой нами величины уровня значимости критерия α (например, $\alpha = 0{,}05$), то гипотезу об отсутствии связи следует отвергнуть, т. е. признать статистическую значимость анализируемой связи. Табл. П.11б построена несколько иначе. В ней при уровне значимости $\alpha = 0{,}05$ и в соответствии с «входами» (m, n) даны «критические» значения величины S, т. е. такие значения, при превышении которых следует отвергать гипотезу об отсутствии связей (признавать их статистическую значимость).

При $n > 7$ для проверки статистической значимости анализируемой связи следует воспользоваться фактом приближенной $\chi^2(n-1)$-распределенности величины $m(n-1) \times \widehat{W}(m)$, справедливым в условиях отсутствия связи в генеральной совокупности ($\widehat{W}(m)$, как и прежде, подсчитывается

по формуле (2.16) или (2.16′)). Поэтому, если окажется, что

$$m(n-1)\,\widehat{W}(m) > \chi^2_\alpha(n-1), \qquad (2.18)$$

то гипотеза об отсутствии ранговой связи между переменными $x^{(k_1)}, x^{(k_2)}, \ldots, x^{(k_m)}$ должна быть отвергнута (с уровнем значимости критерия, равным α); в (2.18) величина $\chi^2_\alpha(n-1)$ — это $100\alpha\%$-ная точка χ^2-распределения с $(n-1)$-й степенью свободы (см. табл. П.4).

Можно использовать и другой способ проверки статистической значимости исследуемой ранговой связи между несколькими переменными, основанный на том, что в условиях отсутствия таковой в генеральной совокупности распределение случайной величины $\frac{1}{2}\ln\frac{(m-1)\cdot\widehat{W}(m)}{1-\widehat{W}(m)}$ приближенно описывается Z-распределением Фишера с числом степеней свободы числителя $\nu_1 = n - 1 - \frac{2}{m}$ и знаменателя $\nu_2 = (m-1)\,\nu_1$ (при большом числе объединенных рангов или значительной их протяженности в расчет ν_1 и ν_2 следует ввести поправку, см. [67, гл. 6]).

Строгих рекомендаций по построению доверительных интервалов для истинного значения W в условиях *наличия* ранговых связей в исследуемой генеральной совокупности к настоящему времени не имеется.

2.3.3. Использование коэффициента конкордации в решении основных задач статистического анализа ранговых связей. Наметим некоторые подходы к решению описанных в п. 2.1.3 задач А, В и С, опирающиеся на понятие коэффициента конкордации.

Задача А. При анализе структуры имеющейся совокупности упорядочений (или структуры связей между исследуемыми порядковыми переменными) существенную пользу может принести решение следующей задачи: найти разбиение анализируемого набора порядковых переменных $x^{(0)}, x^{(1)}, \ldots, \ldots, x^{(p)}$ на заданное число t непересекающихся групп, *оптимальное* в смысле максимизации критерия $\overline{W}(t) = \frac{1}{t}[\widehat{W}_1 + \widehat{W}_2 + \ldots + \widehat{W}_t]$, где $\widehat{W}_j$ — коэффициент конкордации, подсчитанный по переменным, входящим в j-ю группу. Задаваясь различными значениями $t = 2, 3, \ldots, t_0$ $(t_0 < p)$ и прослеживая характер изменения $\overline{W}^{(t)}_{\text{опт}}$ в зависимости от t, можно добиться успеха в выявлении групп высококоррелированных переменных

Задача B. В приложениях, особенно при статистическом анализе совокупности экспертных мнений (представленных в виде ранжировок), существенным оказывается вопрос упорядочения *самих переменных* (интерпретируемых, например, в качестве экспертов) по степени их коррелированности со всеми остальными переменными или с какой-то их частью (представляющей, например, основное ядро высококоррелированных переменных). Для решения этой задачи может быть предложена следующая процедура.

Пусть $\widehat{W}(p+1-k|x^{(j_1)}\ x^{(j_2)}\ \ldots\ x^{(j_k)})$ — коэффициент конкордации, подсчитанный по всем рассматриваемым переменным $x^{(0)}, x^{(1)}, \ldots, x^{(p)}$ за исключением переменных $x^{(j_1)}, \ldots, x^{(j_k)}$. Варьируя состав группы исключенных переменных, мы получим C_{p+1}^k различных значений $\widehat{W}(p+1-k)$. Последовательно вычислим значения всех этих коэффициентов для $k = 0, 1, 2, \ldots, k_0$ и упорядочим их (при каждом фиксированном k) в соответствии с убыванием их значений. Получим:

$\widehat{W}(p+1)$;

$\widehat{W}(p|x^{(j_1)}) \geqslant W(p|x^{(j_2)}) \geqslant \ldots \geqslant W(p|x^{(j_{p+1})})$;

$\widehat{W}(p-1\,|\,x^{(q_1)}, x^{(l_1)}) \geqslant \widehat{W}(p-1\,|\,x^{(q_2)}, x^{(l_2)}) \geqslant \ldots \geqslant W(p-1|x^{(q_L)}, x^{(l_L)})$; $L = C_{p+1}^2$.

. . . .

Эти упорядочения (на каждом «этаже») и дают нам одновременно ранжировки самих переменных (по одной, по паре, по тройке и т. д.) по степени их согласованности с остальными переменными: очевидно, ту переменную (или ту пару, тройку и т. д. переменных), выбрасывание которой приводит к максимальному значению меры согласованности по остальным переменным, естественно объявить наименее связанной (согласующейся) с остальными переменными. Это правило, в частности, было с успехом использовано при обработке экспертных мнений в работе, описанной в [11, § 5.1].

Задача C. Если коэффициент $\widehat{W}(m)$ свидетельствует о наличии статистически значимой связи между анализируемыми показателями $x^{(k_1)}, x^{(k_2)}, \ldots, x^{(k_m)}$, то представляет интерес задача построения оценки неизвестной «истинной» упорядоченности $\widehat{X}^{\text{ист}}$ рассматриваемых объектов. Эта оценка должна быть, по-видимому, результатом некоторого агре-

гирования имеющихся ранжировок $X^{(k_1)}$, $X^{(k_2)}$, ..., $X^{(k_m)}$. Для формирования $\widehat{X}^{(\text{ист})}$ чаще других используют один из трех следующих приемов:

а) компоненты $\widehat{X}^{(\text{ист})}$ определяются в результате сравнения *сумм рангов*, приписываемых каждому объекту упорядочениями $X^{(k_1)}$, $X^{(k_2)}$, ..., $X^{(k_m)}$.

б) компоненты $\widehat{X}^{(\text{ист})}$ определяются в результате сравнения *выборочных медиан рангов*, приписываемых каждому объекту анализируемыми упорядочениями;

в) «присуждение» мест объектам в упорядочении $\widehat{X}^{(\text{ист})}$ основано на «*большинстве голосов*», поданных за данный объект в ранжировках $X^{(k_1)}$, ..., $X^{(k_m)}$ за то или иное место; например, больше других первых мест в анализируемых ранжировках получил объект O_5, тогда ему и присуждается ранг 1 в ранжировке $\widehat{X}^{(\text{ист})}$ и т. д.

2.3.4. Примеры. Рассмотрим примеры, в которых реализуются приведенные выше рекомендации по статистическому анализу множественных ранговых связей.

Пример 2.3. Рассмотрим три порядковые переменные ($x^{(1)}$, $x^{(2)}$, $x^{(3)}$) и соответствующие им упорядочения десяти объектов:

$X^{(1)\prime}$	1	4,5	2	4,5	3	7,5	6	9	7,5	10
$X^{(2)\prime}$	2,5	1	2,5	4,5	4,5	8	9	6,5	10	6,5
$X^{(3)\prime}$	2	1	4,5	4,5	4,5	4,5	8	8	8	10
Сумма	5,5	6,5	9	13,5	12	20	23	23,5	25,5	26,5

В соответствии с формулами (2.18), (2.4) имеем:

$$S=\sum_{i=1}^{10}\left(\sum_{j=1}^{3}x_i^{(j)}-\frac{3\cdot 11}{2}\right)^2=(-11)^2+(-10)^2+(-7,5)^2+$$

$$+(-3)^2+(-4,5)^2+(3,5)^2+(6,5)^2+7^2+9^2+10^2=591;$$

$$T^{(1)}=\frac{1}{12}\cdot(2^3-2)\cdot 2=1;$$

$$T^{(2)}=\frac{1}{12}\cdot(2^3-2)\cdot 3=1,5;$$

$$T^{(3)}=\frac{1}{12}\cdot(4^3-4+3^3-3)=7.$$

Следовательно, в соответствии с (2.16′)

$$\widehat{W}(3)=\frac{591}{\frac{1}{12}\,3^2\cdot(10^2-10)-3\,(1+1{,}5+7)}=\frac{591}{742{,}5-28{,}5}=0{,}828.$$

Пример 2.4. Требуется проверить статистическую значимость множественной ранговой связи 28 переменных ($m = 28$), характеризуемой величиной выборочного коэффициента конкордации $\widehat{W}$ (28) = 0,08, подсчитанного по 13 объектам ($n = 13$).

Воспользуемся фактом χ^2 (12)-распределенности случайной величины $m\,(n-1)\,\widehat{W}\,(m)$, который имеет место (приближенно) в случае, если в исследуемой генеральной совокупности множественная ранговая связь отсутствует. Тогда критерий сводится к проверке неравенства (2.18). Задавшись уровнем значимости критерия $\alpha = 0{,}05$, находим из табл. П.4 значение 5%-ной точки χ^2-распределения с 12 степенями свободы $\chi^2_{0,05}$ (12) = 21,026. В то же время $m\,(n-1)\,\widehat{W}\,(m) = 28\cdot 12\cdot 0{,}08 = 27$.

Поскольку $m\,(n-1)\,\widehat{W}\,(m) > \chi^2_{0,05}\,(12)$, то оказалось, что *даже такого маленького числа, как 0,08*, «хватило» для того, чтобы объявить связь между 28 исследуемыми переменными статистически значимой.

ВЫВОДЫ

1. Анализ статистических связей между порядковыми переменными сводится к статистическому анализу различных упорядочений (ранжировок) одного и того же конечного множества объектов и осуществляется с помощью методов *ранговой корреляции*. В зависимости от типа изучаемой ситуации (шкала измерения анализируемого свойства не известна исследователю или отсутствует вовсе; существуют косвенные или частные количественные показатели, в соответствии со значениями которых можно определять место каждого объекта в общем ряду всех объектов, упорядоченных по анализируемому основному свойству) процесс упорядочения объектов производится либо с привлечением экспертов, либо формализованно — с помощью перехода от исходного ряда наблюдений косвенного *количественного* признака к соответствующему вариационному ряду.

2. Исходные статистические данные для проведения рангового корреляционного анализа представлены *таблицей* (*матрицей*) *рангов* статистически обследованных объектов разме-

ра $n \times (p+1)$ (число объектов на число анализируемых переменных). При формировании матрицы рангов допускаются случаи неразличимости двух или нескольких объектов по изучаемому свойству («объединенные» ранги).

3. К *основным задачам* теории и практики ранговой корреляции относятся: анализ структуры исследуемой совокупности упорядочений (задача А); анализ интегральной (совокупной) согласованности рассматриваемых переменных и их условная ранжировка по критерию степени тесноты связи каждой из них со всеми остальными переменными (задача В); построение единого группового упорядочения объектов на основе имеющейся совокупности согласованных упорядочений (задача С).

4. Статистический анализ взаимосвязей порядковых переменных строится на базе *различных вариантов моделей вероятностного пространства*, в котором роль пространства элементарных исходов играет множество всех возможных перестановок из n элементов (n — число статистически обследованных объектов).

5. В качестве основных характеристик парной статистической связи между упорядочениями используются *ранговые коэффициенты корреляции Спирмэна* $\tau^{(S)}$ *и Кендалла* $\tau^{(K)}$ (см. формулы (2.3)—(2.8)). Значения этих коэффициентов меняются в диапазоне от -1 до $+1$, причем экстремальные значения характеризуют связь соответственно пары прямо противоположных и пары совпадающих упорядочений, а нулевое значение рангового коэффициента корреляции получается при полном отсутствии статистической связи между анализируемыми порядковыми переменными.

6. В качестве основной характеристики статистической связи между *несколькими* (m) порядковыми переменными используется так называемый *коэффициент конкордации* (*согласованности*) *Кендалла* W (m), определяемый формулами (2.16)—(2.16′). Между значением этого коэффициента и значениями парных ранговых коэффициентов Спирмэна, построенных для каждой пары анализируемых переменных, существуют простые соотношения (см. (2.17), (2.17′)).

7. Если представить себе, что каждому объекту некоторой достаточно большой гипотетической совокупности (будем называть ее *генеральной совокупностью*) приписан какой-то ранг по каждой из рассматриваемых переменных и что статистическому обследованию подлежит *лишь часть* этих объектов (*выборка объема* n), то достоверность и практическая ценность выводов, основанных на анализе ранговой корреляции, существенно зависят от следующего вопроса: как ведут себя выборочные значения интересующих нас ранговых корреляционных характе-

ристик при повторениях выборок заданного объема, извлеченных из этой генеральной совокупности. Это и составляет *предмет исследования статистических свойств выборочных ранговых характеристик связи*. Результаты этого исследования относятся прежде всего к построению правил проверки статистической значимости анализируемой связи и к построению доверительных интервалов для неизвестных значений коэффициентов связи, характеризующих всю генеральную совокупность (см. п. 2.2.4, 2.3.2).

8. Парные и множественные характеристики ранговой корреляции являются удобным инструментом решения основных задач (см. задачи А, В и С в п. 2.1.3) статистического анализа связей между порядковыми переменными (см. п. 2.3.3 и примеры 2.1—2.4).

Глава 3. АНАЛИЗ СВЯЗЕЙ МЕЖДУ КЛАССИФИКАЦИОННЫМИ (НОМИНАЛЬНЫМИ) ПЕРЕМЕННЫМИ

3.1. Таблицы сопряженности

Ограничимся рассмотрением таблиц с двусторонней группировкой. Для них сформулированы основные гипотезы и указаны методы их проверки, описана логарифмически-линейная параметризация, приведены различные меры зависимости между строками и столбцами таблицы. Вводятся понятия энтропии случайной величины и информации, содержащейся в одной случайной величине относительно другой случайной величины, представляющие самостоятельный интерес.

Методы изучения таблиц с тремя и более входами можно найти в [23, 75, 154, 168, 199, 238].

3.1.1. Три основные выборочные схемы, приводящие к таблицам сопряженности. *Схема I* возникает в случае, когда распределения строк $(x_{i1}, \ldots, x_{iJ})$ $i = 1, \ldots, I$ (столбцов) таблицы можно рассматривать как независимые выборки из полиномиальных распределений с вероятностями q_{ij}, $\sum_{j=1}^{J} q_{ij} = 1$ и фиксированным числом наблюдений $n_i = \sum_j x_{ij}$. Такая организация данных обычно возникает, когда хотят сравнить между собою несколько одномерных распределений, представленных выборками заранее заданного объема. Наиболее важная гипотеза для первой схемы

$$H_0^1 : q_{ij} = q_{.j}/I, \text{ где } q_{.j} = \sum_i q_{ij}. \quad (3.1)$$

Гипотезу H_0^I называют *гипотезой однородности* (см. [14, п. 1.1.3 и 11.2]).

Схема II. Предполагается, что $(x_{11}, x_{12}, \ldots, x_{IJ})$ имеют полиномиальное распределение с вероятностями $(p_{11}, \ldots, p_{IJ})$ и фиксированным числом наблюдений $n = \sum_{i,j} x_{ij}$. Таблица сопряженности в этом случае является обычной двумерной гистограммой для n наблюдений, а аналогом (3.1) — гипотеза

$$H_0^{II} : p_{ij} = p_{i.}\, p_{.j}, \tag{3.2}$$

где $p_{i.} = \sum_j p_{ij}$ и $p_{.j} = \sum_i p_{ij}$. Если воспользоваться определением условной вероятности [14, п. 4.1.3], то P {попасть в клетку (i, j)| быть в ряду i} $= P$ {быть в столбце j}. Поэтому гипотезу H_0^{II} называют *гипотезой независимости*.

Схема III возникает, когда в схеме II общее число наблюдений рассматривается как случайная величина. Ее важным частным случаем является случай, когда n имеет распределение Пуассона. В этом случае все x_{ij} независимы между собою и также имеют распределение Пуассона с параметрами λ_{ij}. Аналогом (3.1), (3.2) является гипотеза

$$H_0^{III} : \lambda_{ij} = \lambda_{i.} \cdot \lambda_{.j} / \lambda_{..}, \tag{3.3}$$

где $\lambda_{i.} = \sum_j \lambda_{ij}$, $\lambda_{.j} = \sum_i \lambda_{ij}$ и $\lambda_{..} = \sum_{i,j} \lambda_{ij}$. Гипотезу H_0^{III} называют мультипликативной пуассоновской моделью, или, короче, *гипотезой мультипликативности*. В качестве примера схемы III может быть рассмотрена следующая задача. Пусть x_{ij} — число дорожно-транспортных происшествий, зарегистрированных в какой-либо местности в i-й день на дорогах j-го типа. Параметры λ_{ij} в этом случае отражают ожидаемое число дорожно-транспортных происшествий. Если использование транспортом дорог разного типа существенно зависит от дня недели, то гипотеза H_0^{III}, вероятно, не верна. Однако она может иметь место, если, например, рассматривать только рабочие дни.

Существует приближенный графический тест для проверки гипотезы H_0^{III} [154]. Он заключается в том, что для каждого $j = 1, \ldots, J$ строится график, в котором по оси абсцисс откладываются точки $x_{i.} = \sum_j x_{ij}$, а по оси ординат — x_{ij}. Если гипотеза H_0^{III} верна, то нанесенные точки должны группироваться вокруг линии, проходящей через начало координат с наклоном $\lambda_{.j}/\lambda_{..}$. Вероятность выхода заданной точки за

пределы $\pm 2\left(x_{i\cdot}\frac{\lambda_{\cdot j}}{\lambda_{\cdot\cdot}}\left(1-\frac{\lambda_{\cdot j}}{\lambda_{\cdot\cdot}}\right)\right)^{1/2}$ не более 0,05. Использование такого графического представления позволяет сразу же локализовать пары (i, j), в которых происходит значимое отклонение от H_0^{III}.

Можно доказать, что если в схеме III зафиксировать $x_{\cdot\cdot} = n$, то она переходит в схему II с $p_{ij} = \lambda_{ij}/\lambda_{\cdot\cdot}$. При этом H_0^{III} переходит в H_0^{II}. Аналогично, если зафиксировать в схеме II суммы x_{ij} по рядам, положив $n_1 = \sum_j x_{1j}, \ldots, n_I = \sum_j x_{Ij}$, то схема II переходит в схему I с $q_{ij} = p_{ij}/p_{i\cdot}$, а H_0^{II} в H_0^{I}. Поэтому следует ожидать, что в математической трактовке схем I, II, III должно быть много общего.

3.1.2. Логарифмически-линейная параметризация таблиц сопряженности. Для любой из описанных выше моделей положим

$$\mu_{ij}^* = \mathrm{E}x_{ij} = \exp\{\theta^{(0)} + \theta_i^{(1)} + \theta_j^{(2)} + \theta_{ij}^{(12)}\}, \tag{3.4}$$

или

$$\mu_{ij} = \ln \mu_{ij}^* = \theta^{(0)} + \theta_i^{(1)} + \theta_j^{(2)} + \theta_{ij}^{(12)} \tag{3.4'}$$

В данном случае мы несколько отступаем от принятых в книге обозначений, так как индексы сверху θ означают векторы, а индексы снизу — их координаты.

Параметры должны удовлетворять ограничениям

$$\sum_j \theta_{ij}^{(12)} = \sum_i \theta_{ij}^{(12)} = \sum_i \theta_i^{(1)} = \sum_j \theta_j^{(2)} = 0. \tag{3.5}$$

Так же, как в дисперсионном анализе (см. § 13.3), величины $\theta_{ij}^{(12)}$ называют взаимодействиями, $\theta_i^{(1)}$ — эффектами строк, $\theta_j^{(2)}$ — эффектами столбцов и $\theta^{(0)}$ — общим эффектом.

При ограничениях (3.5) модель (3.4) имеет ровно IJ независимых параметров, так как всего имеется одно значение $\theta^{(0)}$, $(I-1)$ независимых $\theta_i^{(1)}$, $(J-1)$ независимых $\theta_j^{(2)}$ и $(I-1)(J-1)$ независимых $\theta_{ij}^{(12)}$.

Из (3.4') и (3.5) следует, что

$$\theta_{ij}^{(12)} = \mu_{ij} - \mu_{i*} - \mu_{*j} + \mu_{**}; \tag{3.6}$$

$$\theta_i^{(1)} = \mu_{i*} - \mu_{**}; \tag{3.7}$$

$$\theta_j^{(2)} = \mu_{*j} - \mu_{**}; \tag{3.8}$$

$$\theta^{(0)} = \mu_{**}, \tag{3.9}$$

где $\mu_{i*} = \sum_j \mu_{ij}/J$, $\mu_{*j} = \sum_i \mu_{ij}/I$, $\mu_{**} = \sum_{i,j} \mu_{i,j}/IJ$.

В новых обозначениях гипотезы H_0^{I}, H_0^{II} и H_0^{III} переходят в гипотезу

$$H_0^{(12)} : \theta_{ij}^{(12)} = 0 \text{ для всех } i \text{ и } j. \tag{3.10}$$

Оценки максимального правдоподобия для параметров получаются из формул (3.6)—(3.9) путем замены в них μ_{ij} на $m_{ij} = \ln x_{ij}$. Если все $x_{ij} \neq 0$, то оценки максимального правдоподобия всегда существуют. Для того чтобы снять проблему существования оценок в общем случае, когда есть $x_{ij} = 0$, положим для всех i, j $m_{ij} = \ln (x_{ij} + c)$, где $0 < c < 1$. Асимптотические (при $n \to \infty$) свойства новых оценок будут такие же, как и у оценок максимального правдоподобия.

3.1.3. Проверка гипотез H_0^{I}, H_0^{II}, H_0^{III}. В [14, п. 11.2.2] описано применение критерия χ^2 для проверки однородности нескольких рядов распределений (гипотеза H_0^{I} в схеме I). В обозначениях настоящего параграфа использованная для этой цели статистика имеет вид

$$X^2 = n \sum_i \sum_j \frac{(x_{ij} - x_{.j} x_{i.}/n)^2}{x_{.j} x_{i.}} = n \left[\sum_i \sum_j \frac{x_{ij}^2}{x_{.j} x_{i.}} - 1 \right]. \tag{3.11}$$

В случае когда H_0^{I} имеет место, X^2 приближенно распределено как χ^2 $((I - 1)(J - 1))$. Этот же критерий можно использовать для проверки гипотез H_0^{II} и H_0^{III}.

Наряду с критерием X^2 для проверки этих гипотез применяют информационную статистику

$$2n\widehat{I} = 2\Big(\sum_i \sum_j x_{.j} \ln x_{ij} - \sum_i x_{i.} \ln x_{i.} -$$
$$- \sum_j xj, \ln x_{.j} + x_{..} \ln x_{..}\Big), \tag{3.12}$$

которая при выполнении гипотезы о независимости или однородности при $n \to \infty$ асимптотически так же распределена, как χ^2 $((I - 1)(J - 1))$. Когда в таблице одна или несколько клеток содержат нули, рекомендуется применять поправку: для каждого нуля отнимать из величины $2n\widehat{I}$ единицу[1]. Критерий $2n\widehat{I}$ легко получить из общих принципов проверки сложной гипотезы [14, п. 9.3.3]. В самом деле,

$$- 2 \ln \{L (x_{11}, \ldots, x_{IJ}; \widetilde{\Theta}^{(1)}, \widetilde{\Theta}^{(2)}, \Theta^{(12)} = 0)/L (x_{11}, \ldots, x_{IJ};$$
$$\widehat{\Theta}^{(1)}, \widehat{\Theta}^{(2)}, \widehat{\Theta}^{(12)})\}$$

---

[1] См.: Закс Л. Статистическое оценивание. — М.: Статистика, 1976, с. 445.

равняется правой части (3.12), а гипотеза $H_0^{(12)}$ накладывает ограничения на $(I-1)(J-1)$ параметров. Большие расхождения между критериями X^2 и $2n\hat{I}$ на практике наблюдаются редко.

Особое значение информационной статистики заключается в том, что для таблиц с многосторонней группировкой она может быть разложена на аддитивные составляющие, соответствующие различным гипотезам. При этом может быть построена теория, во многом параллельная дисперсионному анализу (см. гл. 13).

3.1.4. Меры связи между строками и столбцами таблицы. Если связь, обнаруживающаяся при проверке гипотез независимости или однородности, оказывается значимой, то полезно иметь численную меру ее. Хотя величина X^2 дает нам удобный критерий значимости связи, она не может служить мерой связи. Так, если оставить неизменными все относительные величины в таблице и увеличивать общее число измерений n, то величина X^2 будет расти пропорционально n. Предложено много различных мер связи [23], но наиболее известными среди них являются меры, основанные на отношении X^2/n:

$\varphi = (X^2/n)^{1/2}$ — *квадратный корень из среднего квадрата сопряженности*;

$C = [X^2/(X^2+n)]^{1/2}$ — *коэффициент сопряженности*;

$T = [X^2/n\sqrt{(I-1)(J-1)}]^{1/2}$ — *мера связи Чупрова*.

Наряду с ними практический интерес представляют информационные меры связи. Прежде чем переходить к ним, введем понятие энтропии случайной величины и информации, содержащейся в одной случайной величине относительно другой случайной величины.

Пусть случайная величина ξ принимает конечное число значений x_i $(i = 1, \ldots, k)$ с вероятностями, соответственно равными $p_\xi(x_i)$, тогда

$$H(\xi) = -\sum_x p_\xi(x) \ln p_\xi(x) = -\mathbf{E} \ln p_\xi(\xi) \tag{3.13}$$

называют *энтропией* ξ и рассматривают в качестве меры неопределенности ξ. Энтропия обладает следующими свойствами:

1) $H(\xi) \geqslant 0$, причем равенство достигается тогда и только тогда, когда ξ принимает только одно значение;

2) $H(\xi)$ не меняется при взаимно-однозначных преобразованиях ξ;

3) $H(\xi)$ максимально, когда все возможные значения ξ равновероятны.

По аналогии с $H(\xi)$ определяются *энтропия* распределения *пары случайных величин* и *условная энтропия*:

$$H(\xi, \eta) = -\sum_{x,y} p_{\xi,\eta}(x, y) \ln p_{\xi,\eta}(x, y) = -\mathbf{E} \ln p_{\xi,\eta}(\xi, \eta);$$

$$H(\xi \mid \eta) = \sum_{y} p_{\eta}(y) H(\xi \mid \eta = y) = -\sum_{y} p_{\eta}(y) \sum_{x} p_{\xi}(x \mid \eta = y) \ln p_{\xi}(x \mid \eta = y).$$

Основные свойства $H(\xi|\eta)$:

1) $H(\xi, \xi) = H(\xi)$;

2) $H(\xi, \eta) = H(\xi) + H(\eta|\xi)$;

3) $H(\xi, \eta) \leqslant H(\xi) + H(\eta)$, причем равенство достигается тогда и только тогда, когда ξ и η статистически независимы.

Основные свойства $H(\xi|\eta)$:

1) $H(\xi|\xi) = 0$;

2) $H(\xi|\eta) \leqslant H(\xi)$, причем равенство достигается тогда и только тогда, когда ξ и η статистически независимы.

Информационная мера зависимости ξ и η определяется как

$$I(\xi, \eta) = H(\xi) + H(\eta) - H(\xi, \eta). \quad (3.14)$$

Про $I(\xi, \eta)$ говорят, что она измеряет количество информации в ξ относительно η или количество информации в η относительно ξ. Основные свойства $I(\xi, \eta)$ легко следуют из свойств $H(\xi, \eta)$:

1) $I(\xi, \eta) \geqslant 0$, причем равенство достигается тогда и только тогда, когда ξ и η статистически независимы;

2) $I(\xi, \xi) = H(\xi)$.

При анализе таблиц сопряженности используют *направленные меры связи*:

$$C_{\xi|\eta} = I(\xi, \eta)/H(\xi); \quad (3.15)$$

$$C_{\eta|\xi} = I(\xi, \eta)/H(\eta). \quad (3.15')$$

Коэффициенты C заключены в пределах между 0 и 1 и по своим свойствам во многом аналогичны обычным коэффициентам корреляции. Они равны нулю, когда переменные ξ и η статистически независимы; $C_{\xi|\eta}$ равно 1, когда ξ однозначно определяется по η; они не меняются при взаимно-однозначных преобразованиях переменных.

Пример 3.1. Пусть взаимное распределение ξ и η задано с помощью табл. 3.1.

Тогда $H(\xi, \eta) = 1{,}06$; $H(\xi) = 0{,}77$; $H(\eta) = 1{,}06$; $I(\xi, \eta) = 0{,}77$; $C_{\xi|\eta} = 1$; $C_{\eta|\xi} = 0{,}73$.

Таблица 3.1

ξ	η			p_ξ
	0	1	2	
0	0,40	0,20	0	0,60
1	0	0	0,40	0,40
p_η	0,40	0,20	0,40	1

3.2. Приписывание численных значений качественным переменным *(дуальное шкалирование)*

3.2.1. Методическое место дуального шкалирования. Наряду со статистическими методами, изложенными в предыдущем параграфе, в работе с таблицами сопряженности может быть использован принципиально отличный подход. Градациям переменных, измеренных в общем случае в шкалах наименований, приписываются численные значения так, чтобы достиг своего экстремума определенный функционал. Далее с новыми переменными работают как с переменными, измеренными в качественных шкалах. В целом этот подход, который мы, следуя предложенному в [232], будем называть *дуальным шкалированием* (dual scaling), по своему методическому содержанию ближе к анализу данных, чем к традиционным статистическим методам. В нем не формулируется математическая модель распределения исходных данных, предлагаемые статистические критерии носят, вообще говоря, эвристический характер, но зато четко и наглядно формулируется принцип приписывания численных значений.

Дуальное шкалирование за последние 50 лет открывалось и переоткрывалось независимо разными исследователями и известно под различными названиями: «метод взаимных усреднений» (the method of reciprocal averages) [210, 246], «аддитивное или оптимальное шкалирование» (additive or optimal scoring) [183], «метод максимизации коэффициента корреляции» (bivariate correlation approach) [198, 230], «взвешивание по Гутману» (Guttman weighting) [169], «анализ главных компонент качественных данных» (principal component analysis

of qualitative data), «одновременная линейная регрессия» (simultaneous linear regression) [203]. С точки зрения используемого алгебраического аппарата к дуальному шкалированию примыкают современные методы визуализации табличных данных: «биплот» (biplot) [171], «(факторный) анализ соответствий» (correspondance (factor) analysis) [165, 166], хотя их целевая направленность шире задачи оцифровки значений переменных.

Широкое использование при обработке данных ЭВМ сделало дуальное шкалирование одним из основных инструментов первичного анализа данных. Этим объясняется возрождение внимания к нему в начале 70-х годов. Основные публикации, последних лет по дуальному шкалированию — [222, 224, 232, 247]. Вычислительные программы могут быть найдены в [169] и [232].

Сопоставление различных подходов к выбору оптимизируемого функционала в дуальном шкалировании позволяет глубже понять заложенные в методе возможности. Поэтому в дальнейшем сформулируем несколько различных принципов приписывания численных значений и покажем, что все они ведут к одному и тому же результату.

3.2.2. Максимизация F-отношения суммы квадратов отклонений между объектами к полной сумме квадратов отклонений. Изложение начнем с гипотетического численного примера. Предположим, что 10 экспертов произвели оценку организации труда в четырех лабораториях. Эксперты могли использовать лишь три категории оценок: хорошо, удовлетворительно, неудовлетворительно, и один из экспертов оценивал лишь первые три лаборатории. Пусть полученные данные представлены в виде таблицы сопряженности X, в которой x_{ij} означает число оценок градации j, полученных i-й лабораторией (табл. 3.2).

Таблица 3.2

Порядковый номер лаборатории (i)	Оценка (j)			Итого ($x_{i\cdot}$)
	хорошо	удовлетворительно	неудовлетворительно	
1	1	3	6	10
2	2	4	4	10
3	3	5	2	10
4	6	3	0	9
Итого $x_{\cdot j}$	12	15	12	$x_{\cdot\cdot}=39$

Припишем численные значения оценкам: v_1 — хорошо, v_2 — удовлетворительно, v_3 — неудовлетворительно. Тогда набранные лабораториями оценки можно представить в виде односторонней таблицы дисперсионного анализа (см. гл. 13), в которой $СК_п$ — полная сумма квадратов отклонений, $СК_м$ — сумма квадратов отклонений между лабораториями и $СК_{вн}$ — сумма квадратов отклонений внутри лабораторий (табл. 3.3):

Таблица 3.3

Порядковый номер лаборатории (i)	Оценки (w_{ij})	Итого ($w_{i.}$)
1	v_1, v_2, v_2, v_2, v_3, v_3, v_3, v_3, v_3, v_3	$v_1+3v_2+6v_3=w_1.$
2	v_1, v_1, v_2, v_2, v_2, v_2, v_3, v_3, v_3, v_3	$2v_1+4v_2+4v_3=w_2.$
3	v_1, v_1, v_1, v_2, v_2, v_2, v_2, v_2, v_3, v_3	$3v_1+5v_2+2v_3=w_3.$
4	v_1, v_1, v_1, v_1, v_1, v_1, v_2, v_2, v_2	$6v_1+3v_2=w_4.$
Итого		$12v_1+15v_2+12v_3=$ $=w_1.+w_2.+w_3.+w_4.=$ $=w_{..}$

0. $x_{..}=12+15+12=39$;
1. $C=w_{..}^2/x_{..}$;
2. $\sum_i \sum_j w_{ij}^2=12v_1^2+15v_2^2+12v_3^2$;
3. $\sum_i (w_{i.}/x_{i.})^2=w_{1.}^2/10+$ $+w_{2.}^2/10+w_{3.}^2/10+w_{4.}^2/9$;
4. $СК_п=(2)-(1)$;
5. $СК_м=(3)-(1)$;
6. $СК_{вн}=(2)-(3)$.

Величину $\eta^2 = СК_м/СК_{вн}$ будем называть *корреляционным отношением* (correlation ratio). Поскольку $СК_п = СК_м + СК_{вн}$, то $0 \leqslant \eta^2 \leqslant 1$.

Если $v_1 = v_2 = v_3 = \text{const}$, то $СК_м = СК_{вн} = СК_п = 0$ и η^2 не определено. Исключим этот неинтересный случай и подберем численные значения v_i так, чтобы оптимизировать η^2. Полагая для однозначности $w_{..} = 0$, $v_1 > 0$, $СК_п = 39$ и временно опуская детали вычислений, имеем

$$\eta^2_{max}=0{,}292;\quad V=\begin{pmatrix}v_1\\ v_2\\ v_3\end{pmatrix}=\begin{bmatrix}1{,}234\\ 0{,}064\\ -1{,}313\end{bmatrix};$$

$$W=\begin{pmatrix}w_1\\ w_2\\ w_3\\ w_4\end{pmatrix}=\begin{pmatrix}w_{1.}/x_{1.}\\ w_{2.}/x_{2.}\\ w_{3.}/x_{3.}\\ w_{4.}/x_{4.}\end{pmatrix}=\begin{bmatrix}-1{,}194\\ -0{,}468\\ 0{,}258\\ 1{,}560\end{bmatrix}. \qquad (3.16)$$

Таким образом, нами одновременно решены следующие задачи: 1) приписаны численные значения $\{v_i\}$ градациям оценок, которыми пользовались эксперты; 2) оценена в условных единицах (баллах) $\{w_j\}$ организация труда в лабораториях; 3) оптимизировано корреляционное отношение (η^2).

Дадим общую формулировку принципа, по которому приписываются численные значения, и опишем соответствующую вычислительную процедуру.

Матричная формулировка основного принципа оптимизации. Пусть $\mathbf{X}$ — $(m \times n)$-матрица таблицы сопряженности; $A = (x_{1.}, \ldots, x_{m.})'$ — $(m \times 1)$-вектор сумм элементов $\mathbf{X}$ по строкам; $B = (x_{.1}, \ldots, x_{.n})'$— $(n \times 1)$-вектор сумм элементов $\mathbf{X}$ по столбцам; $x_{..}$ — общая сумма элементов $\mathbf{X}$;

$$\underset{(m\times m)}{\mathbf{W}}=\begin{bmatrix}x_{1.} & & 0\\ & \ddots & \\ 0 & & x_{m.}\end{bmatrix};\quad \underset{(n\times n)}{\mathbf{V}}=\begin{bmatrix}x_{.1} & & 0\\ & \ddots & \\ 0 & & x_{.n}\end{bmatrix}\text{—}$$

вспомогательные диагональные матрицы; $V = (v_1, \ldots, v_n)'$ — $(n \times 1)$-вектор численных значений, приписанных строкам; $W = (w_1, \ldots, w_m)'$ — $(m \times 1)$-вектор численных значений, приписанных столбцам; $\bar{w} = \sum_i w_i x_{i.}/x_{..} = A'W/x_{..} = B'V/x_{..}$ — общее среднее значение. В нашем примере $m = 4$, $n = 3$, $\mathbf{X}$ определяется по табл. 3.2; $A = (10, 10, 10, 9)'$; $B = (12, 15, 12)'$; $x_{..} = 39$;

$$\mathbf{W}=\begin{pmatrix}10 & 0 & 0 & 0\\ 0 & 10 & 0 & 0\\ 0 & 0 & 10 & 0\\ 0 & 0 & 0 & 9\end{pmatrix};\quad \mathbf{V}=\begin{pmatrix}12 & 0 & 0\\ 0 & 15 & 0\\ 0 & 0 & 12\end{pmatrix};$$

векторы V, W определяются из (3.16); $\bar{w} = 0$.

При сделанных предположениях (ср. с табл. 3.3)

$$СК_п = V'\,\mathbf{V}V,\quad СК_м = V'\,\mathbf{X}'\,\mathbf{W}^{-1}\mathbf{X}V, \qquad (3.17)$$

откуда

$$\eta^2 = V' \mathbf{X}' \mathbf{W}^{-1} \mathbf{X} V / V' \mathbf{V} V \qquad (3.18)$$

при условии, что

$$B'V = V'B = 0. \qquad (3.19)$$

Оптимизация величины η^2. Поскольку уравнениями (3.18) и (3.19) V определяется с точностью до постоянного множителя, положим для определенности

$$V'\mathbf{V}V = x\ldots \qquad (3.20)$$

Будем искать максимум числителя (3.18) при ограничениях (3.19) и (3.20) методом множителей Лагранжа. Пусть $Q(V) = V'\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}V - \lambda_1 (V'\mathbf{V}V - x_{..}) - \lambda_2 V'B$, тогда для нахождения V должны быть решены уравнения

$$\partial Q/\partial V = 2\,[\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}]\,V - 2\lambda_1 \mathbf{V}V - \lambda_2 B = 0; \qquad (3.21)$$

$$\partial Q/\partial \lambda_1 = V'\mathbf{V}V - x_{..} = 0; \qquad (3.22)$$

$$\partial Q/\partial \lambda_2 = V'B = 0. \qquad (3.23)$$

Умножим (3.21) слева на V' и, воспользовавшись уравнением (3.23), получаем с учетом (3.18), что
$\lambda_1 = V'\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}V/V'\mathbf{V}V = \eta^2$.

Для оценки величины λ_2 умножим (3.21) слева на $\mathbf{1}_n = (1, \ldots, 1)'$ и воспользуемся легко проверяемыми равенствами

$$\mathbf{1}_n'\mathbf{X}'\mathbf{W}^{-1}\mathbf{X} = A'\mathbf{W}^{-1}\mathbf{X} = \mathbf{1}_m'\mathbf{W}^{-1}\mathbf{W}\mathbf{X} = \mathbf{1}_m'\mathbf{X} = B';$$
$$\mathbf{1}_n'\mathbf{V} = B'. \qquad (3.24)$$

В силу (3.19) отсюда следует, что $\lambda_2 = 0$. Уравнения (3.21) могут теперь быть представлены в виде

$$(\mathbf{V}^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{X} - \eta^2\,\mathbf{I})\,V = 0; \qquad (3.25)$$

таким образом, η^2 должно быть собственным значением уравнения (3.25). Поскольку легче работать с симметричными матрицами, произведем замену переменных, положив

$$\tilde{\mathbf{V}} = \mathbf{V}^{1/2}V. \qquad (3.26)$$

Уравнения (3.21)—(3.23) при этом перепишутся в виде

$$(\mathbf{V}^{-1/2}\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\mathbf{V}^{-1/2} - \eta^2\,\mathbf{I})\,\tilde{V} = 0; \qquad (3.21')$$

$$\tilde{\mathbf{V}}'\,\tilde{V} = x_{..}; \qquad (3.22')$$

$$\tilde{\mathbf{V}}'\,\mathbf{V}^{1/2}\,\mathbf{1}_n = 0. \qquad (3.23')$$

По аналогии с цепочкой уравнений (3.24) непосредственно, проверяется, что вектор $\widetilde{V}_0 = \mathbf{V}^{1/2}\mathbf{1}_n$ является собственным вектором (3.21′), отвечающим собственному числу $\eta^2 = 1$ удовлетворяет (3.22′) и не удовлетворяет (3.23′). Отсюда следует, что искомое η^2 будет вторым по порядку после 1 собственным числом (3.21′), а вектор $\widetilde{V}$ — соответствующим ему собственным вектором. При этом будет выполнено и условие (3.23′), так как собственные векторы, отвечающие разным числам, взаимно перпендикулярны.

С помощью стандартной алгебраической процедуры [102, гл. 5] можно исключить из матрицы $\mathbf{R} = \mathbf{V}^{-1/2}\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\mathbf{V}^{-1/2}$ собственное число $\eta^2 = 1$. Для этого $\mathbf{R}$ достаточно заменить на

$$\mathbf{C} = \mathbf{R} - \mathbf{V}^{1/2}\mathbf{1}_n \cdot \mathbf{1}_n' \mathbf{V}^{1/2}/x_{..}. \tag{3.27}$$

Нахождение максимального собственного числа и соответствующего ему собственного вектора уравнения $(\mathbf{C} - \eta^2\mathbf{I}) \times \times \widetilde{V} = 0$ проводится стандартными методами [102, гл. 4].

3.2.3. Двойственность в определении V **и** W. В примере предыдущего пункта мы приписывали веса различным градациям оценок и получили баллы для лабораторий. При этом

$$\eta^2 = \sup_{\widetilde{V} \neq \mathbf{V}^{1/2}\mathbf{1}_n} \frac{\widetilde{V}'\mathbf{V}^{-1/2}\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\mathbf{V}^{-1/2}\widetilde{V}}{\widetilde{V}'\widetilde{V}} \tag{3.28}$$

Однако аналогичную задачу мы могли бы решить в обратном порядке, приписывая численные веса лабораториям так, чтобы максимизировать среднеквадратический разброс между средними баллами, соответствующими разным градациям оценок. При этом по аналогии мы получили бы

$$\eta^2 = \sup_{\widetilde{W} \neq \mathbf{W}^{1/2}\mathbf{1}_m} \frac{\widetilde{W}'\mathbf{W}^{-1/2}\mathbf{X}\mathbf{V}^{-1}\mathbf{X}'\mathbf{W}^{-1/2}\widetilde{W}}{\widetilde{W}'\widetilde{W}}. \tag{3.29}$$

Если обозначить $\mathbf{D} = \mathbf{V}^{-1/2}\mathbf{X}'\mathbf{W}^{-1/2}$, то в формуле (3.28) η^2 является вторым по величине собственным числом матрицы $\mathbf{DD}'$, а в (3.29) — вторым по величине собственным числом $\mathbf{D}'\mathbf{D}$. Известно [102], что отличные от нуля собственные числа матриц $\mathbf{DD}'$ и $\mathbf{D}'\mathbf{D}$ совпадают. Следовательно, совпадают и значения η^2.

Возьмем теперь вектор V_1, вычисленный как решение (3.21)—(3.23), и найдем вектор W_1 средних значений, приписанных лабораториям

$$W_1 = \mathbf{W}^{-1}\mathbf{X}V_1. \tag{3.30}$$

Будем считать, что лабораториям приписаны численные значения, определяемые вектором W_1, и вычислим средние баллы, которые получат градации оценок:

$$V_2 = \mathbf{V}^{-1}\mathbf{X}'W_1 = \mathbf{V}^{-1}\mathbf{X}'\,\mathbf{W}^{-1}\mathbf{X}V_1. \tag{3.31}$$

В силу (3.25) левая часть (3.31) есть $\eta^2 V_1$, т. е. векторы V_2 и V_1 пропорциональны.

Продолжая, мы находим

$$W_2 = \mathbf{W}^{-1}\mathbf{X}V_2 = \mathbf{W}^{-1}\mathbf{X}\eta^2 V_1 = \eta^2 W_1. \tag{3.32}$$

Таким образом, безразлично, решим ли мы экстремальную задачу п. 3.2.2 для столбцов или строк, мы определим значения V, W с точностью до множителя пропорциональности.

На этом свойстве, а также на том, что η^2 — максимальное собственное число, меньшее 1, основан метод взаимных усреднений. В нем выбирается значение V_1 (или W_1) так, чтобы выполнялось соотношение (3.23), далее по формулам (3.30) и (3.31) находятся W_1 и V_2. Вектор V_2 каким-либо образом нормируется, например умножается на величину, обратную максимальному абсолютному значению его координат. Процесс вычислений повторяется до тех пор, пока последовательные значения V не будут близки друг к другу. Условие (3.23) гарантирует, что у начального вектора V_1 нет составляющей, соответствующей $\eta^2 = 1$. Описанный итерационный процесс сходится тем быстрее, чем удачнее выбрано начальное приближение.

3.2.4. Максимизация коэффициента корреляции. Рассматривая матрицу $\mathbf{X}$ в качестве выборки из двумерного распределения (V, W) и для простоты выкладок полагая $B'V = A'W = 0$, можно определить коэффициент корреляции между переменными как

$$r = \frac{W'\mathbf{X}V}{\sqrt{W'\,\mathbf{W}W \cdot V'\,\mathbf{V}V}}. \tag{3.33}$$

Будем теперь V и W искать из условия максимизации значения r. Для этого, так же как в 3.2.2, воспользуемся методом множителей Лагранжа.

Пусть $Q(V, W) = W'\mathbf{X}V - \lambda_1 (V'\mathbf{V}V - x_{..}) - \lambda_2 (W' \times \times \mathbf{W}W - x_{..})$, тогда уравнения для нахождения V и W имеют вид:

$$\partial Q/\partial V = \mathbf{X}'\,W - 2\lambda_1 \mathbf{V}V = 0; \tag{3.34}$$

$$\partial Q/\partial W = \mathbf{X}V - 2\lambda_2 \mathbf{W}W = 0; \tag{3.35}$$

$$\partial Q/\partial\lambda_1 = V'\,\mathbf{V}V - x_{..} = 0; \qquad (3.36)$$

$$\partial Q/\partial\lambda_2 = W'\,\mathbf{W}W - x_{..} = 0. \qquad (3.37)$$

Умножив слева (3.34) на V', а (3.35) на W' и воспользовавшись (3.36), (3.37), (3.33), имеем

$$V'\,\mathbf{X}'\,W = 2\lambda_1 V'\,\mathbf{V}V = 2\lambda_1 x_{..};$$

$$W'\,\mathbf{X}V = 2\lambda_2 W'\,\mathbf{W}W = 2\lambda_2 x_{..};$$

$$W'\,\mathbf{X}V = rx_{..}.$$

Откуда $r = 2\lambda_1 = 2\lambda_2$. Воспользовавшись (3.35), (3.36), (3.37), заменим в уравнении для определения r W через V:

$$r = \max_{W,\,V} \frac{W'\,\mathbf{X}V}{\sqrt{W'\,\mathbf{W}W V'\,\mathbf{V}V}} = \frac{1}{r} \max_{V} \frac{V'\,\mathbf{X}'\,\mathbf{W}^{-1}\mathbf{X}V}{V'\,\mathbf{V}V} = \frac{\eta^2}{r}.$$

Таким образом, $r^2 = \eta^2$ и V является собственным вектором (3.25), т. е. максимизация коэффициента корреляции приводит к тем же численным значениям, что и изложенные выше методы.

3.2.5. Изучение оптимального решения. Когда найдено оптимальное решение (η^2, V, W), возникает вопрос, в какой степени оно исчерпывает информацию, содержащуюся в исходных данных. Ведь у матрицы **C** (см. (3.27)) есть другие собственные значения и векторы. По аналогии с методом главных компонент [14, § 10.5] для ответа на этот вопрос будем использовать величину

$$\delta = 100\% \cdot \eta^2/\mathrm{Sp}\,(\mathbf{C}), \qquad (3.38)$$

где Sp (**C**) — сумма диагональных элементов **C**.

В примере п. 3.2.2 $\delta = 92{,}3\%$, т. е. оптимальное решение (3.21′)—(3.23′) объясняет существенную долю информации, содержащейся в табл. 3.2.

Строго обоснованного теста для проверки значимости отличия от нуля оптимального значения η^2 нет. В [232] рекомендуется приближенный критерий

$$\chi^2 = -[x_{..} - 1 - (n + m - 1)/2]\ \ln\,(1 - \eta^2) \qquad (3.39)$$

с числом степеней свободы $f = m + n - 3$. В нашем случае $f = 4 + 3 - 3 = 4$,

$$\chi^2 = -[39 - 1 - (3 + 4 - 1)/2]\ \ln\,(1 - 0{,}292) = 12{,}10.$$

Различие следует считать значимым с уровнем значимости 0,017.

Для сравнения к тем же данным применим статистические критерии из § 3.1.

Традиционный критерий (3.11): $X^2 = 12{,}35$ ($f = 6$); уровень значимости связи между переменными — 0,055.

Логлинейный подход (3.12): $2n\widehat{I} = 13{,}13$ (с поправкой на нулевую ячейку), ($f = 6$); уровень значимости связи между переменными — 0,041.

Из приведенных данных видно, что в рассмотренном примере с точки зрения оценки статистической значимости связи между строками и столбцами традиционный и логлинейный подходы к таблицам сопряженности, с одной стороны, и дуальное шкалирование, с другой стороны, дают сравнительно близкие результаты. Однако в общем случае связь между этими двумя методами пока достаточно не изучена [232, с. 181].

3.2.6. Таблицы «объект — многомерный отклик». Исходные данные для дву-, трех- и более мерных таблиц сопряженности часто могут быть представлены в форме таблицы, в которой строки соответствуют объектам (субъектам), столбцы — градациям используемых классификационных переменных и на пересечении i-й строки и столбца, соответствующего j-й градации l-й переменной, стоит 1 или 0 в зависимости от того, имеет ли место для i-го объекта эта градация (1) или нет (0). В случае когда для ряда объектов значения одной из переменных не определены (измерены в непредусмотренной шкале, не измерены, утрачены при обработке и т. п.), либо исключают из таблицы соответствующие объекты, либо вводят для этой переменной дополнительную градацию «значение не определено». Пример фрагмента таблицы, которая могла бы быть исходной для данных примера п. 3.2.2, дан в табл. 3.4, где приведена оценка организации труда в четырех лабораториях (таблица «Единица наблюдения —(лаборатория, оценка, эксперт)»). В качестве единицы наблюдения (объекта) в ней взято резюме из карточки, заполняемой экспертом после обследования и оценки организации труда в лаборатории, в котором указываются номер лаборатории, оценка, номер эксперта.

Если бы априори было известно, что эксперты эквивалентны друг другу и их оценки не зависят как от обследуемых ими лабораторий, так и от оценок, выставляемых другими исследователями, то в табл. 3.2 и 3.4 содержалась бы одна и та же информация, но представленная в разном виде. Однако на практике эти условия обычно не выполняются. Так из табл. 3.4 видно, что первый эксперт по сравнению со вторым имеет тенденцию завышать оценки. Поэтому табл. 3.4 содержит больше информации, чем табл. 3.2, и дает возможность изучить не

Таблица 3.4

Единица наблюдения	Лаборатория				Оценка*				Эксперт										Итого
	1	2	3	4	1	2	3	4	1	2	3	4	5	6	7	8	9	10	
1	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	3
2	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	3
3	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	3
4	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	3
5	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	3
6	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	3
7	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	3
8	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
37	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	3
38	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	3
39	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	3
40	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	3
Итого	10	10	10	10	12	15	12	1	4	4	4	4	4	4	4	4	4	4	120

* Градации оценок: хорошо, удовлетворительно, неудовлетворительно, не определено.

только соотношение между лабораториями и экспертами (связь экспертов с лабораториями), но и соотношение между экспертами и оценками (средний критический уровень эксперта).

Приписывание численных значений в таблицах «Объект — (многомерный отклик)» можно провести по полной аналогии с тем, как это сделано в п. 3.2.2 [232]. Введем необходимые обозначения, указывая в скобках соответствующий аналог в обозначениях п. 3.2.2:

l — число переменных в отклике;

n_k — число градаций k-й переменной;

$n = \sum n_k$ — общее число градаций переменных (n);

m — число объектов (субъектов, единиц наблюдения) (m);

$\mathbf{F}$ — ($m \times n$) -матрица данных, состоящая из нулей и единиц ($\mathbf{X}$);

F — ($n \times 1$)-вектор сумм по столбцам (A);

G — ($m \times 1$) -вектор сумм по строкам (B);

$\mathbf{D}$ — ($n \times n$)-диагональная матрица сумм по столбцам ($\mathbf{V}$);

lI_m — диагональная матрица сумм по строкам ($\mathbf{W}$);

$f_{..}$ — общая сумма элементов $\mathbf{F} = l \cdot m$ ($x_{..}$);

X — ($n \times 1$) -вектор значений, приписываемых переменным (V);

Y — ($m \times 1$)-вектор значений, приписываемых объектам W).

Условия (3.19) теперь формулируются так, что сумма взвешенных откликов внутри переменных должна равняться нулю. С учетом этого условия $СК_п = СК_м + СК_{ост} + СК_{пер}$, где $СК_{пер} = 0$ — сумма квадратов отклонений между средними значениями по переменным, $СК_п$ — полная сумма квадратов отклонений в таблице, $СК_м$ — сумма квадратов отклонений между строками, $СК_{ост}$ — остаточная сумма квадратов отклонений. Оптимизации подвергается величина $\lambda^2 = СК_м/СК_п$.

В случае таблиц «Объект — (многомерный отклик)» так же сохраняется свойство взаимозаменяемости оптимизации, выполняемой по строкам, и оптимизации по столбцам. При выполнении условий $F'X = G'Y = 0$ и с учетом более простого вида аналога $\mathbf{W}$ имеем

$$\lambda^2 = \frac{X'\,\mathbf{F}'\,\mathbf{F}X/l}{X'\,\mathbf{D}X} = \frac{Y'\,\mathbf{F}\mathbf{D}^{-1}\,\mathbf{F}'\,Y}{lY'\,Y}.$$

Если оптимизация проводится по X, то

$$\mathbf{C} = l^{-1}\cdot\mathbf{D}^{-1/2}\,\mathbf{F}'\,\mathbf{F}\mathbf{D}^{-1/2} - f_{..}^{-1}\,\mathbf{D}^{1/2}\,\mathbf{1}_n\,\mathbf{1}'_n\,\mathbf{D}^{1/2};$$

если по Y, то

$$\mathbf{C}_1 = l^{-1}\,\mathbf{F}\mathbf{D}^{-1}\,\mathbf{F}' - m^{-1}\,\mathbf{1}_n\,\mathbf{1}'_n.$$

Если дуальное шкалирование выполнить для таблицы сопряженности и соответствующей ей таблицы «Объект — (двумерный отклик)», то численные значения, приписанные переменным после нормировки, совпадут, а $\lambda^2 = (1 + \eta)/2$.

ВЫВОДЫ

1. Распределения многомерных случайных величин, координаты которых измеряются в номинальных и порядковых шкалах, часто представляют в виде многомерных прямоугольных таблиц, называемых таблицами сопряженности. При этом в клетке, соответствующей i_1 — градации первой переменной, ..., i_k — k-й переменной указывается $x_{i_1 \ldots i_k}$ — число наблюдений в выборке с этими градациями. В двумерном случае по организации сбора данных различают три выборочные схемы, приводящие к таблице сопряженности:

1) распределения столбцов (строк) независимы и являются полиномиальными распределениями с вероятностями $\{q_{ij}\}$ и фиксированным числом наблюдений в столбце $n_i = \sum_j x_{ij}$; основная гипотеза: $\{q_{ij}\}$ от i не зависит;

2) распределение частот в двумерной таблице есть полиномиальное распределение с вероятностями $\{p_{ij}\}$ и фиксированным числом наблюдений $n = \sum_{i,j} n_{ij}$; основная гипотеза: для всех i, j $p_{ij} = p_{i.}p_{.j}$.

3) все x_{ij} независимы между собою и имеют пуассоновское распределение с параметрами λ_{ij}; основная гипотеза в этом случае: для всех i, j $\lambda_{ij} = \lambda_{i.}\lambda_{.j}/\lambda_{..}$.

2. Для описания совместного распределения x_{ij} предложена логарифмически-линейная параметризация таблиц сопряженности, в которой предполагается, что

$$\ln \mathrm{E} x_{ij} = \theta^{(0)} + \theta_i^{(1)} + \theta_j^{(2)} + \theta_{ij}^{(12)},$$

где параметры θ удовлетворяют соотношениям

$$\sum_i \theta_i^{(1)} = \sum_j \theta_j^{(2)} = \sum_i \theta_{ij}^{(12)} = \sum_j \theta_{ij}^{(12)} = 0.$$

В новых обозначениях основная гипотеза записывается как

$$H_0^{(12)}: \text{для всех } i, j \ \theta_{ij}^{(12)} = 0.$$

Для проверки этой гипотезы используются либо обычный критерий X^2 как критерий однородности распределений столбцов (строк) в таблице, либо имеющий то же асимптотическое распределение информационный критерий $2n\widehat{I}$, получаемый стандартным способом для проверки сложной гипотезы $\theta_{ij}^{(12)} = 0$.

3. Предложен ряд различных мер связи между строками и столбцами в двумерных таблицах сопряженности. Среди них выделяются информационные меры связи как легко допускающие обобщение на многомерный случай.

4. Один из методов анализа двумерных таблиц сопряженности заключается в том, чтобы приписать градациям классификационных переменных численные значения так, чтобы максимизировать некоторый функционал. Оказывается, что ряд известных под различными названиями и максимизирующих различные функционалы методов таких, как «метод взаимных усреднений», «аддитивное или оптимальное шкалирование», «метод максимизации коэффициента корреляции» и др., приводит к приписыванию одних и тех же численных значений.

Глава 4. АНАЛИЗ СТРУКТУРЫ СВЯЗЕЙ МЕЖДУ КОМПОНЕНТАМИ МНОГОМЕРНОГО ВЕКТОРА

4.1. Связи прямые и опосредованные. Введение в проблематику

4.1.1. Цепи Маркова. Рассмотрим такую последовательность случайных (для определенности непрерывных) величин

$$\xi_1, \xi_2, \ldots, \xi_n, \ldots, \tag{4.1}$$

что для каждого $k = 2, \ldots, n$ условное распределение ξ_k при $\xi_{k-1} = x_{k-1}$ совпадает с условным распределением ξ_k при условии, что $\xi_{k-1} = x_{k-1}$, $\xi_{k-2} = x_{k-2}$, ..., $\xi_1 = x_1$ для всех наборов $x_{k-1}, \ldots, x_1$, для которых соответствующие условные распределения определены. На языке условных плотностей это условие может быть записано так: для всех $k \geqslant 2$

$$f_{\xi_k}(x_k \mid \xi_{k-1} = x_{k-1}) = f_{\xi_k}(x_k \mid \xi_{k-1} = x_{k-1}, \ldots, \xi_1 = x_1) \tag{4.2}$$

или, для краткости опуская значения случайных величин и нижние индексы у буквы f, для всех $k \geqslant 2$

$$f(\xi_k|\xi_{k-1}) = f(\xi_k|\xi_{k-1}, \ldots, \xi_1). \tag{4.2$'$}$$

Про последовательность (4.1) говорят, что она образует *цепь Маркова*. В цепи Маркова каждый член зависит от всех предшествующих, но непосредственно зависящими (связанными) можно в силу (4.2) считать только члены, стоящие рядом, рассматривая не рядом стоящие члены ξ_l и ξ_{l+k} $(k \geqslant 2)$ как связанные опосредованно через $\xi_{l+1}, \ldots, \xi_{l+k-1}$.

Пример 4.1. Пусть случайные величины $\eta_1, \ldots, \eta_n$ независимы между собой и нормально распределены со средним 0 и дисперсией 1, ρ — некоторая константа $0 < \rho < 1$, а $\xi_1 = \eta_1$,

$$\xi_2 = \rho\xi_1 + (1-\rho^2)^{1/2}\eta_2, \ \xi_3 = \rho\xi_2 + (1-\rho^2)^{1/2}\eta_3, \ \ldots$$

Тогда случайные величины $\xi_1, \xi_2, \ldots, \xi_n$ также имеют нормальное распределение с теми же параметрами, что и η_i, и связаны в цепь Маркова. Их корреляционная матрица имеет вид

$$\begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{n-3} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{bmatrix},$$

т. е. зависит всего от одного параметра ρ. Непосредственно связанные члены имеют коэффициент корреляции ρ, а члены, опосредованно связанные и отстоящие друг от друга на k членов последовательности, имеют меньший коэффициент корреляции ρ^{k+1}. Таким образом, чем связь непосредственнее, тем она сильнее.

Удобна геометрическая иллюстрация цепи Маркова, при которой случайные величины изображаются точками или кружками, а непосредственные (прямые) связи между ними — соединяющими их отрезками (рис. 4.1). Для обозначения связей мы использовали отрезки, а не стрелки, так как если последовательность (4.1) образует цепь Маркова, то и последовательность $\xi_n, \xi_{n-1}, \ldots, \xi_1$, как нетрудно убедиться (см., например, [78, с. 590—591]), также является цепью Маркова. Цепи Маркова являются простейшей моделью зависимостей между случайными величинами и нашли очень широкое применение в практике (физика, техника, экономика, биология, лингвистика) особенно в тех случаях, когда есть естественное (например, временное) упорядочение случайных величин [26, 62, 96].

①—②—③—···—ⓝ—···

Рис. 4.1. Прямые связи между случайными величинами, образующими цепь Маркова

В кратких обозначениях формулы (4.2′) плотность совместного распределения $\xi_1, \ldots, \xi_n$ может быть выражена как

$$f(\xi_1, \ldots, \xi_n) = f(\xi_1) f(\xi_2 \mid \xi_1) f(\xi_3 \mid \xi_2, \xi_1) \ldots \cdot f(\xi_n \mid \xi_{n-1}, \ldots, \xi_1) = f(\xi_1) f(\xi_2 \mid \xi_1) \ldots f(\xi_n \mid \xi_{n-1}). \quad (4.3)$$

Откуда следует, что для описания распределения цепи Маркова достаточно знать распределение первого члена последовательности и для $i = 2, \ldots, n$ — условные распределения ξ_i при известном значении ξ_{i-1}, т. е. плотности условных распределений пар векторов, непосредственно связанных друг с другом. Это свойство используется ниже при введении понятий прямой и опосредованной связи между координатами вектора.

4.1.2. Прямые связи между координатами вектора. По аналогии с первым равенством формулы (4.3) по формуле условной вероятности для координат p-мерного вектора $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})'$, имеющего невырожденное непрерывное распределение, имеем

$$f(\xi^{(1)}, \ldots, \xi^{(p)}) = f(\xi^{(1)}) f(\xi^{(2)} \mid \xi^{(1)}) f(\xi^{(3)} \mid \xi^{(2)}, \xi^{(1)}) \ldots \ldots f(\xi^{(p)} \mid \xi^{(p-1)}, \ldots, \xi^{(1)}). \quad (4.4)$$

Предположим теперь, что для каждого $i = 2, \ldots, p$ найдется такое $j(i) < i$, что выражение в правой части (4.4) может быть представлено в форме, близкой к правой части (4.3), а именно

$$f(\xi^{(1)}, \ldots, \xi^{(p)}) = f(\xi^{(1)})\, f(\xi^{(2)} \mid \xi^{(j(2))}) \ldots f(\xi^{(p)} \mid \xi^{(j(p))}). \qquad (4.5)$$

В этом случае пары координат с номерами $(2, j(2))$, $(3, j(3))$, ..., $(p, j(p))$ можно назвать непосредственно (прямо) связанными, а остальные координаты считать связанными опосредованно. В общем случае естественно отказаться от ограничений, накладываемых нумерацией координат вектора, предполагая, что существует такая перестановка индексов координат, при которой представление вида (4.5) возможно. Удобно также ввести значение $j = 0$ как соответствующее неслучайной дополнительной координате $\xi^{(0)} = 1$.

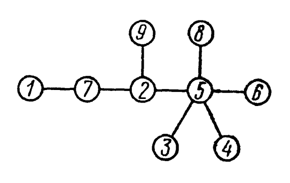


Рис. 4.2. Прямые связи, выделенные при изучении структуры трудовых ресурсов

Пример 4.2 [150]. На рис. 4.2 графически показаны прямые связи, выделенные при изучении структуры трудовых ресурсов. Рассматривалась 9-мерная случайная величина, реализациями которой являлись значения показателей по 71 региону РСФСР за 1969 г. Использовались следующие показатели: 1) доля среднегодовой численности рабочих, служащих, колхозников в среднегодовой численности населения; 2) доля специалистов с высшим и средним специальным образованием, занятых в народном хозяйстве, в среднегодовой численности рабочих, служащих, колхозников; 3) доля специалистов с высшим и средним специальным образованием, занятых в сельском хозяйстве, в общей численности работающих в сельском хозяйстве; 4) доля работающих в промышленности и строительстве в среднегодовой численности рабочих, служащих, колхозников; 5) доля работающих в сельском хозяйстве в среднегодовой численности рабочих, служащих, колхозников; 6) доля работающих на транспорте и в связи в среднегодовой численности рабочих, служащих, колхозников; 7) доля работающих в области просвещения, науки, культуры, искусства в среднегодовой численности рабочих, служащих, колхозников; 8) доля работающих в области государственного и хозяйственного управления, кредита, государственного страхования в среднегодовой численности рабочих, служащих, колхозников; 9) доля работающих в области здравоохранения, физической

культуры, социального обеспечения в среднегодовой численности рабочих, служащих, колхозников.

На рис. 4.2 хорошо видна на изучаемый год центральная роль в распределении трудовых ресурсов по отраслям народного хозяйства доли занятых в сельском хозяйстве (показатель *5*). Это хорошо согласуется с качественными представлениями специалистов по трудовым ресурсам. Обращает на себя внимание тесная связь показателей *2* и *9*, что также допускает качественное истолкование.

Предположением (4.5) введен новый малопараметрический класс распределений, обобщающий многомерные распределения, которые возникают в цепях Маркова, и получивший название *«распределения с древообразной структурой зависимостей»* (ДСЗ). Происхождение этого названия будет ясно из материала следующего параграфа, где в более строгой и полной форме даны все необходимые определения и рассмотрены свойства нормальных распределений с ДСЗ. Можно ожидать, что в приложениях новый класс распределений окажется столь же удобным инструментом, каким сегодня являются цепи Маркова при изучении временных рядов. Первые результаты использования распределений с ДСЗ очень обнадеживают [113].

Распределения с ДСЗ были введены в статистическую практику С. Чоу [174, 175, 176]. Если не считать краткого изложения результатов Чоу в [48], они не нашли еще отражения в монографической литературе. В отечественной литературе разработка теоретических вопросов, примыкающих к этому новому направлению, дана в [40, 61]. На работы В. И. Заруцкого [58, 59] мы существенно опираемся в последующем изложении.

4.1.3. Математические задачи, связанные с изучением распределений с ДСЗ. Прежде всего надо более четко описать класс распределений с ДСЗ и выявить соотношения между различными параметризациями одного и того же распределения, возникающими при разном упорядочении координат. Ведь даже в простейшем случае, когда координаты образуют цепь Маркова, возможны два упорядочения: в прямом направлении цепи Маркова и в обратном. Необходимо также найти аналог выявленному на цепях Маркова соотношению, что прямым связям отвечает более высокая корреляция между координатами (см. § 4.2).

Нужно научиться оценивать структуру связей по выборочным данным. Было бы желательно исследовать свойства этой процедуры как в обычной асимптотике растущего объема, так и в специальной более адекватной для многомерных данных асимптотике, когда рассматривается последовательность задач восстановления структуры зависимостей, в которой при пере-

ходе от одной задачи к другой одновременно растут и объем выборки, и число координат вектора (см. § 4.3).

Если несколько видоизменить формулу (4.3), оставив под знаком условия не один предшествующий член, а $m \geqslant 2$ предшествующих, т. е.

$$f(\xi_1, \ldots, \xi_n) = f(\xi_1) f(\xi_2 \mid \xi_1) f(\xi_3 \mid \xi_2, \xi_1) \ldots$$
$$\ldots f(\xi_n \mid \xi_{n-1}, \ldots, \xi_{n-m}), \quad (4.6)$$

то придем к так называемым *m-зависимым Марковским цепям*. Естественно понятие m-зависимости перенести на координаты вектора (см. § 4.4).

При изучении связей между координатами мы уже использовали геометрический язык, изображая координаты точками, а связи между ними — соединяющими их отрезками. Это язык теории графов. Терминология и методы теории графов широко используют при изложении основного материала этой главы. Поэтому ниже приводятся предварительные сведения из теории графов.

4.2. Распределение с древообразной структурой зависимостей

4.2.1. Предварительные сведения из теории графов. Изложение начнем с напоминания основных понятий теории графов [134].

Определение 4.1. *Простым графом* G называется пара $(V(G), E(G))$, где $V(G)$ — непустое конечное множество элементов, называемых *вершинами графа* G ($V(G)$ — множество вершин G), а $E(G)$ — конечное множество неупорядоченных пар различных элементов из $V(G)$, называемых *ребрами графа* G ($E(G)$ — множество ребер G). В дальнейшем термин «простой» опускается. Отметим, что так как $E(G)$ определено как множество, а не как совокупность и состоит из неупорядоченных элементов, то в графе G каждую пару вершин $a, b \in V(G)$ может соединять не более чем одно ребро (a, b) и $(a, b) = (b, a)$. В дальнейшем (как и на рис.4.1 и 4.2) вершины графа мы будем отождествлять с координатами вектора, а ребра графа — со связями.

Определение 4.2. Граф G_1 называется подграфом G, если $V(G_1) \subset V(G)$ и $E(G_1) \subset E(G)$.

Определение 4.3. Конечная непустая последовательность ребер графа G $M = \{(a_1, a_2), (a_2, a_3), \ldots, (a_m, a_{m+1})\}$ называется *простой цепью*, соединяющей вершины a_1 и a_{m+1},

если все вершины $a_1, \ldots, a_{m+1}$ различны, кроме, быть может, $a_{m+1} = a_1$. В последнем случае простая цепь называется *циклом*.

Определение 4.4. Граф G называется *связанным*, если для любых его вершин a и b существует простая цепь, соединяющая a и b.

Определение 4.5. *Лесом* называется граф, не содержащий циклов, связанный лес называется *деревом*.

Графы, изображенные на рис. 4.1 и 4.2, связанные и не имеют циклов. Следовательно, их можно назвать деревьями. На них легко проверяются утверждения следующей теоремы.

Теорема 4.1 [134]. *Определяющие свойства графа-дерева.* Пусть граф T имеет p вершин, тогда следующие утверждения эквивалентны: 1) T является деревом; 2) T не содержит циклов и имеет $(p - 1)$ ребер; 3) T связан и имеет $(p - 1)$ ребер; 4) любые две вершины T соединены ровно одной простой цепью; 5) T не содержит циклов, но, добавляя к нему любое новое ребро, мы получим ровно один цикл.

4.2.2. Распределения с древообразной структурой зависимостей (ДСЗ). Изложение начнем с определения.

Определение 4.6. Будем говорить, что p-мерный вектор X имеет ДСЗ, если существует хотя бы одна перестановка координат вектора $\alpha(1, \ldots, p) = (\alpha(1), \alpha(2), \ldots, \alpha(p))$, такая, что для каждого $\alpha(i)$ найдется номер

$$j(\alpha(i)) \in \{0, \alpha(1), \ldots, \alpha(i-1)\}, \tag{4.7}$$

что «почти всюду по $\{x^{(\alpha(1))}, \ldots, x^{(\alpha(p))}\}$» [1] для всех z

$$P\{x^{(\alpha(i))} < z \mid x^{(\alpha(1))}, \ldots, x^{(\alpha(i-1))}\} = P\{x^{(\alpha(i))} < z \mid x^{(j(\alpha(i)))}\}. \tag{4.8}$$

При этом $j = 0$ соответствует фиктивной координате $x^{(0)} \equiv 1$ и $j(\alpha(1)) = 0$.

Для вектора X с ДСЗ рассмотрим граф $G = (V, E)$, где $V = \{0, 1, \ldots, p\}$ и $E = \bigcup_i (i, j(i))$. Граф G имеет p ребер и в силу (4.7) не имеет цикла, поэтому согласно п.2 теоремы 4.1 он является деревом. Отсюда и происходит термин *«древообразная структура зависимостей»*. Граф G будем называть *графом структуры зависимостей* X. Заметим, что в случае, когда для некоторого $\alpha(i)$ можно положить $j(\alpha(i)) = 0$, т. е. распределение $x^{(\alpha(i))}$ не зависит от $x^{(\alpha(1))}, \ldots, x^{(\alpha(i-1))}$, то за $j(\alpha(i))$

---

[1] В дальнейшем изложении взятые в кавычки слова, выражающие сугубо внутриматематическое требование общности, опускаются.

можно было бы выбрать любое из чисел, стоящих в правой части (4.7). Таким образом, граф G определяется, вообще говоря, неоднозначно.

Однако единственность будет, если на распределение X наложить дополнительное ограничение; для всех пар координат $x^{(i)}$, $x^{(j)}$, для всех возможных значений $x^{(i)} = u$ и $x^{(j)} = v$ в случае дискретного распределения X

$$P\{x^{(i)} = u,\ x^{(j)} = v\} > 0 \tag{4.8}$$

и в непрерывном случае ($i,\ j \geqslant 1$)

$$f_{x^{(i)},\ x^{(j)}}(u,\ v) > 0. \tag{4.8'}$$

В важном частном случае невырожденного p-мерного нормального распределения условие (4.8′) выполняется всегда.

Теорема 4.2. Пусть вектор X имеет ДСЗ, выполняются условия (4.8) и (4.8′) и G_1 и G_2 — два различных графа структуры зависимостей X. Тогда для любого ребра $(i,\ j) \in E\ (G_1)$ и $\notin E\ (G_2)$, координаты (вектора) $x^{(i)}$ и $x^{(j)}$ независимы, т. е. графы G_1 и G_2 отличаются друг от друга только ребрами, соответствующими независимым координатам. Ввиду принципиальной важности этого результата изложим схему его доказательства. Оно проводится в несколько шагов.

1. В графе G_2 выбирается простая цепь, соединяющая вершины i и j. Согласно п. 4 теоремы 4.1 она всегда существует. Так как $(i,\ j) \notin E\ (G_2)$, цепь содержит хотя бы одну вершину, отличную от i, j. Обозначим эту вершину l.

2. Координаты $x^{(i)}$, $x^{(l)}$, $x^{(j)}$, как лежащие на простой цепи (в графе G_2), образуют марковскую последовательность. Следовательно, в дискретном случае совместное распределение $x^{(i)}$, $x^{(l)}$, $x^{(j)}$ описывается формулой

$$P(x^{(i)})\,P(x^{(l)}|x^{(i)})\,P(x^{(j)}|x^{(l)}). \tag{4.9}$$

3. Возьмем в графе G_2 простую цепь, соединяющую l и i. Возможны два случая: 1) цепь содержит вершину j и 2) цепь не содержит вершину j. В первом случае на простой цепи вершины лежат в порядке l, j, i; во втором — в порядке l, i, j. Оба случая рассматриваются одинаково. Пусть для определенности имеет место первый случай, тогда совместное распределение $x^{(l)}$, $x^{(j)}$, $x^{(i)}$ описывается формулой

$$P(x^{(l)})\,P(x^{(j)}\,|\,x^{(l)})\,P(x^{(i)}\,|\,x^{(j)}). \tag{4.10}$$

4. Формулы (4.9) и (4.10) описывают одно и то же распределение, поэтому их можно приравнять. Опираясь на условие

(4.8), в полученном равенстве можно произвести упрощения. После несложных преобразований получаем

$$P(x^{(j)})\, P(x^{(l)}, x^{(i)}) = P(x^{(i)}, x^{(j)})\, P(x^{(l)}).$$

Произведем суммирование по всем возможным значениям $x^{(l)}$. В результате получаем, что

$$P(x^{(i)}, x^{(j)}) = P(x^{(i)})\, P(x^{(j)}),$$

что и требовалось доказать. Случай непрерывных распределений рассматривается аналогично с заменой вероятностей на соответствующие плотности.

Рассмотрим теперь задачу о нахождении при известном графе структуры зависимостей G перестановки координат α, позволяющей представить распределение X в виде (4.5). Положим $\alpha(0) = 0$ и возьмем произвольную простую цепь, начинающуюся в 0. Будем двигаться вдоль нее от нуля, считывая номера проходимых координат и приравнивая их $\alpha(1)$, $\alpha(2)$, ... Затем берем следующую простую цепь, начинающуюся в одной из уже пройденных вершин или в 0, и двигаемся вдоль нее, продолжая считывание, и т. д. до тех пор, пока не будут исчерпаны все вершины графа и тем самым определена полностью перестановка α. Поскольку координаты, лежащие вдоль простой цепи, образуют цепь Маркова (см. п. 2, 3 схемы доказательства теоремы 4.2), из построения α сразу же следует возможность представления распределения X в виде (4.5). В отдельных случаях перед построением α может оказаться удобным в графе G изменить некоторые несущественные связи, соответствующие независимым координатам (ср. с теоремой 4.2).

4.2.3. Нормальное распределение с ДСЗ. Пусть X имеет невырожденное p-мерное распределение с вектором средних M и ковариационной матрицей $\Sigma = ||\sigma_{ij}||$ с известной структурой зависимостей, заданной функцией $j(i)$. Вопросы, связанные с нахождением $j(i)$, обсуждаются в следующем параграфе. Наша ближайшая цель — найти общий вид плотности X.

Известно (см. [14, с. 172] и теорему 2.5.1 [20, с. 45]), что условное распределение $x^{(i)}$ при фиксированном значении компоненты $x^{(j)}$ нормально с параметрами

$$\mathbf{E}(x^{(i)} \mid x^{(j)}) = m^{(i)} + r_{i,j}\frac{\sigma_i}{\sigma_j}(x^{(j)} - m^{(j)});$$

$$\mathbf{D}(x^{(i)} \mid x^{(j)}) = \sigma_i^2(1 - r_{ij}^2),$$

где $\sigma_i^2=\sigma_{ii}$, $r_{ij}=\sigma_{ij}/\sigma_i\sigma_j$. Откуда в силу (4.5) плотность X равна:

$$(2\pi)^{-p/2}\prod_{1\leqslant i\leqslant p}\sigma_i^{-1}(1-r^2_{ij\,(i)})^{-1/2}\times$$
$$\times\exp\left\{-\frac{\left(x^{(i)}-m^{(i)}-\dfrac{\sigma_{ij\,(i)}}{\sigma^2_{j\,(i)}}\left(x^{(j\,(i))}-m^{(j\,(i))}\right)\right)^2}{2\sigma_i^2\left(1-r^2_{ij\,(i)}\right)}\right\}. \qquad (4.11)$$

Таким образом, гауссовские распределения с ДСЗ имеют очень простой вид Σ^{-1} — матрицы, обратной ковариационной. В ней над диагональю стоят не более $p-1$ отличных от нуля элементов. Если перестановка α совпадает с исходной нумерацией координат X, то над главной диагональю в каждом столбце Σ^{-1} стоит не более одного отличного от нуля элемента.

В качестве примера приведем ковариационные матрицы случайных векторов, графы структуры зависимостей которых показаны на рис. 4.1 и 4.2. В первом случае

$$\Sigma^{-1}=(1-\rho^2)^{-n}\begin{bmatrix} 1 & -\rho & 0\ldots & 0 & 0\\ -\rho & 1 & -\rho\ldots & 0 & 0\\ 0 & -\rho & 1\ldots & 0 & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & 0\ldots & 1 & -\rho\\ 0 & 0 & 0\ldots & -\rho & 1 \end{bmatrix}$$

и во втором

$$\Sigma^{-1}=\begin{bmatrix} * & 0 & 0 & 0 & 0 & 0 & * & 0 & 0\\ 0 & * & 0 & 0 & * & 0 & * & 0 & *\\ 0 & 0 & * & 0 & * & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & * & * & 0 & 0 & 0 & 0\\ 0 & * & * & * & * & * & 0 & * & 0\\ 0 & 0 & 0 & 0 & * & * & 0 & 0 & 0\\ * & * & 0 & 0 & 0 & 0 & * & 0 & 0\\ 0 & 0 & 0 & 0 & * & 0 & 0 & * & 0\\ 0 & * & 0 & 0 & 0 & 0 & 0 & 0 & * \end{bmatrix},$$

здесь знаком * показаны отличные от нуля элементы.

Полезно представление Σ^{-1} в виде

$$\Sigma^{-1}=C'\,C, \qquad (4.12)$$

где $\mathbf{C} = \|c_{il}\|$ — матрица с элементами

$$c_{ii} = \sigma_i^{-1}(1 - r_{ij(i)}^2)^{-1/2}; \quad c_{ij(i)} = -r_{ij(i)}\sigma_{j(i)}^{-1}(1 - r_{ij(i)}^2)^{-1/2};$$
$$c_{il} = 0, \ l \neq i, \ l \neq j(i). \tag{4.13}$$

Если перестановка α совпадает с исходной нумерацией координат, то $j(i) < i$ и $\mathbf{C}$ — нижняя треугольная матрица.

Граф структуры зависимостей G нормально распределенного вектора X может быть использован при вычислении коэффициентов корреляции между координатами X. Для этого нам необходимо знать только p коэффициентов корреляции между парами координат, соответствующих ребрам G.

Теорема 4.3. Для нормального вектора X с ДСЗ для всех $1 \leqslant i < j \leqslant p$

$$r_{ij} = \prod_{k,\, l\, :\, (k,\, l) \in M(i,\, j)} r_{kl}, \tag{4.14}$$

где $M(i, j)$ — простая цепь, связывающая в графе G структуры зависимостей вершины i и j.

Доказательство. Последовательность координат X, образующая простую цепь, является марковской (см. п.2 и 3 доказательства теоремы 4.2). Пусть эти координаты будут i, l_1, l_2, ..., l_k, j. В силу теоремы 1 [140, с. 122] для последовательности нормальных величин, связанных в цепь Маркова,

$$r_{ij} = r_{il_1} \cdot r_{l_1 j} = r_{il_1} \cdot r_{l_1 l_2} \cdot r_{l_2 j} = \ldots = r_{il_1} r_{l_1 l_2} \ldots r_{l_k j},$$

что и требовалось доказать.

Остановимся теперь на выборочной оценке Σ^{-1} при известном графе G структуры зависимостей. В качестве первого шага по графу G находится перестановка α. Это можно, например, сделать так, как указано в конце предыдущего пункта. Далее строится $\widehat{\mathbf{C}}$ — оценка матрицы $\mathbf{C}$ путем замены в $\mathbf{C}$ величин σ_i, $r_{ij(i)}$ их выборочными оценками. $\mathbf{S}^{-1}$ — оценка Σ^{-1} находится как

$$\mathbf{S}^{-1} = \widehat{\mathbf{C}}' \widehat{\mathbf{C}}. \tag{4.12'}$$

Если в качестве $\widehat{\sigma_i}$, $\widehat{r_{ij(i)}}$ взять обычные в нормальном случае выборочные оценки [14, табл. 6.3, п. 6], то $\left(\frac{n-1}{n}\mathbf{S}\right)^{-1}$ есть оценка максимального правдоподобия [14, § 8.2] для Σ^{-1} при известной структуре зависимостей. Для доказательства этого можно воспользоваться леммой 3.2.2 [20], позволяющей найти в рассматриваемом случае максимум уравнения правдоподобия.

4.3. Оценка графа структуры зависимостей компонент нормального вектора

4.3.1. Вес связи. Пусть X — нормально распределенный вектор с ДСЗ своих компонент. *Весом связи* (i, j) назовем $|r_{ij}|$, где r_{ij} — коэффициент корреляции между $x^{(i)}$ и $x^{(j)}$. *Весом графа* назовем суммарный вес его ребер. Тогда вес графа структуры зависимостей

$$w(G) = \sum_{(i,\, j(i)) \in E(G)} |r_{ij(i)}|. \tag{4.15}$$

Формула (4.14) подсказывает, что среди всех деревьев T, которые можно построить на вершинах $\{0, 1, \ldots, p\}$, графы структуры зависимостей, отличающиеся между собой (в силу теоремы 4.2) только несущественными связями с нулевым весом, будут иметь наибольший вес.

Теорема 4.4. Для невырожденного нормального вектора с ДСЗ вес графа структуры зависимостей строго больше веса любого дерева, отличающегося от него хотя бы одним ребром, имеющим ненулевой вес.

Доказательство теоремы проводится методом математической индукции. Для $p = 2$ оно верно. Предположим, что оно верно для всех $p' \leqslant p$, и докажем, что оно верно и для $p + 1$. Не нарушая общности, можно считать, что перестановка α соответствует естественной нумерации координат. Обозначим G_p и T_p граф структуры зависимостей X и произвольное дерево, построенные на $V(G_p) = \{0, 1, 2, \ldots, p\}$. Тогда $E(G_{p+1}) = E(G_p) + (p + 1,\ j(p + 1))$ и $E(T_{p+1}) = E(T_p + (p + 1, k)$, где $k < p + 1$ — некоторая вершина G. Если $k = j(p + 1)$, то утверждение теоремы верно согласно предположению, так как

$$w(G_{p+1}) = w(G_p) + |r_{p+1,\, j(p+1)}|; \tag{4.16}$$
$$w(T_{p+1}) = w(T_p) + |r_{p+1,\, k}|.$$

Но если $k \neq j(p+1)$, то согласно (4.14) $|r_{p+1,\, k}| =$

$$= |r_{p+1,\, j(p+1)}| \cdot |r_{j(p+1),\, l_1}| \cdots |r_{l_i,\, k}|,$$

где вершины $p + 1,\ j(p + 1),\ l_1, \ldots, l_i,\ k$ берутся вдоль простой цепи в G_{p+1}, соединяющей $p + 1$ и k. Поскольку все коэффициенты корреляции по модулю строго меньше единицы (в силу невырожденности распределения X), то

$$|r_{p+1,\, k}| < |r_{p+1,\, j(p+1)}|, \tag{4.17}$$

если только $|r_{p+1,\ j(p+1)}| \neq 0$. В силу сделанного предположения, а также (4.16) и (4.17) утверждение теоремы верно и для $p+1$.

4.3.2. Построение графа структуры зависимостей по корреляционной матрице. Как установлено выше, граф G структуры зависимостей нормального вектора строго тяжелее любого дерева, построенного на тех же вершинах и отличающегося от G хотя бы одним ребром ненулевого веса. Поэтому задача нахождения G при известной корреляционной матрице $\mathbf{R} = ||r_{ij}||$ сводится к задаче отыскания среди деревьев, которые можно построить на вершинах $V(G)$ с весами, определяемыми $\mathbf{W} = |||r_{ij}|||$, дерева наибольшего веса. В теории графов последняя задача решается с помощью алгоритма Крускала [134], носящего итерационный характер и заключающегося в следующем:

сначала матрица $\mathbf{W}$ пополняется весами, отвечающими 0—координате $x^{(0)} \equiv 1$, $\mathbf{W}^0 = \begin{bmatrix} 0\ 0 \ldots 0 \\ 0 & \\ \ldots & \\ 0 & \mathbf{W} \end{bmatrix}$;

далее в качестве первого шага выбирается любое из ребер, имеющих в $\mathbf{W}^0$ наибольший вес; на l-м шаге $(2 \leqslant l \leqslant p)$ — любое из ребер наивысшего веса среди оставшихся и не образующих цикла с ранее выбранными ребрами.

Поскольку всего имеется $p+1$ вершина, в алгоритме Крускала делается p шагов, и на каждом из них выбирается ребро, не образующее цикла с ранее выбранными, то в результате его применения возникает дерево (см. теорему 4.1). Работу алгоритма Крускала удобно проиллюстрировать на примере построения дерева для однородной цепи Маркова, описанной в п. 4.1.1. На каждом из первых $n-1$ шагов выбираются ребра вида $(i, i+1)$, на последнем шаге — ребро вида $(0, j)$, так как все остальные ребра образуют цикл с ранее выбранными. Если отбросить связь нулевого веса, то получаем дерево, изображенное на рис. 4.1.

Если известна только выборочная корреляционная матрица $\widehat{\mathbf{R}}$, то по ней может быть построена выборочная весовая функция $\mathbf{W} = |||\widehat{r_{ij}}|||$. Результат применения к ней алгоритма Крускала обозначим $\widehat{G}$. Так как при росте объема выборки $\widehat{\mathbf{R}} \to \mathbf{R}$ (по вероятности), то $\widehat{G}$ также сходится к G в том смысле, что

$$w\,(E\,(\widehat{G} \setminus G) \cup E\,(G \setminus \widehat{G})) \to 0 \text{ (по вероятности)}^1. \qquad (4.18)$$

---

[1]Здесь $\widehat{G} \setminus G$ означает множество элементов $\widehat{G}$, не входящих в G.

4.3.3. Асимптотика Колмогорова — Деева. В практической работе часто p — размерность вектора X и n — число наблюдений суть величины одного порядка.

Например, в медицинских исследованиях при диагностике относительно редких заболеваний приходится работать с векторами размерности $p = 10—15$ при выборках объема $n = 20—30$. Ясно, что в этих условиях результаты типа (4.18), установленные в предположении, что распределение фиксировано, а $n \to \infty$, вряд ли могут служить надежным обоснованием.

В последние годы получила распространение новая асимптотика, специально рассчитанная на многомерные задачи, в которых отношение p/n не стремится к нулю. В этой асимптотике рассматривается последовательность (по некоторому параметру $m \to \infty$) многомерных задач изучаемого класса. При росте m (переходе в последовательности от одной задачи к другой) растут как $p(m)$, так и $n(m)$, причем их отношение стремится к пределу

$$p(m) \to \infty, \; n(m) \to \infty, \; p(m)/n(m) \to \lambda < \infty \; (m \to \infty). \qquad (4.19)$$

В этой специальной асимптотике, которую мы в дальнейшем будем называть асимптотикой Колмогорова — Деева, нарушаются многие привычные свойства статистических процедур. Например, если X имеет многомерное нормальное распределение с нулевым вектором средних и независимыми координатами с дисперсией σ^2 и X_i $(i = 1, \ldots, n)$ — независимая выборка объема n, то квадрат длины вектора выборочного среднего

$$\sum_{k=1}^{p} (\bar{x}^{(k)})^2 = \sum_{k=1}^{p} \left(\sum_{i=1}^{n} x_i^{(k)}/n \right)^2 \to \lambda\sigma^2,$$

а не к 0, как это было бы в обычной асимптотике.

Достоинство новой асимптотики не в том, что в ней не обязательно верны многие общепринятые статистические процедуры, а в том, что полученные в ней предельные формулы, например для ошибок классификации многомерных объектов, исключительно хорошо работают даже при относительно небольших значениях n.

Алгоритм Крускала оказывается устойчивым по отношению к новой асимптотике. Так, если равномерно по m для некоторого $\delta > 0$

$$\begin{aligned} &1 - \max_{i,j} |r_{ij}| > n^{-1/2+\delta}; \\ &\min_{i,j;\, r_{ij} \neq 0} |r_{ij}| > 2n^{-1/2+\delta}, \end{aligned} \qquad (4.20)$$

т. е. при переходе от одной задачи к другой в асимптотике Колмогорова — Деева $\max |r_{ij}|$ по всем парам координат не приближается слишком быстро к единице, а $\min |r_{ij}|$ по существенным (ненулевого веса) связям не стремится слишком быстро к нулю, то (4.18) имеет место и в асимптотике (4.19). При этом выборочные значения коэффициентов корреляции совсем не обязаны удовлетворять соотношению (4.14), задающему свойство древообразности для нормальных распределений. Они только должны быть близки к теоретическим значениям коэффициентов, которые удовлетворяют (4.14).

4.4. $R(k)$-распределения

Распределения с ДСЗ обобщают совместное распределение последовательных членов в дискретных цепях Маркова. Если двигаться вдоль ветвей графа-дерева структуры зависимостей, то последовательно проходимые вершины графа (координаты вектора наблюдений) образуют цепь Маркова. Этот факт позволил доказать единственность в нормальном случае графа — дерева структуры зависимостей, предложить простой алгоритм его оценки по выборочной корреляционной матрице и, наконец, показать, как, зная дерево структуры зависимостей, получить исходное распределение.

В этом параграфе рассматриваются $R(k)$-распределения (удовлетворяющие условию $R(k)$), обобщающие так называемые k-зависимые марковские последовательности. $R(1)$-распределение — это уже известное нам распределение с ДСЗ. По аналогии со случаем $k = 1$ вводится понятие графа структуры зависимостей и показывается, как найти этот граф по выборочным данным. Однако в общем случае ($k > 1$) пока не удалось доказать однозначность обратного перехода: восстановления по графу структуры зависимостей и $(k+1)$-мерным распределениям координат вектора X исходного распределения X. Этим обусловлена некоторая незавершенность излагаемой ниже теории.

4.4.1. Основные определения. Начнем с обобщения понятия распределения с ДСЗ.

Определение 4.7. Распределение X удовлетворяет условию $R(k)$ $(k \geqslant 1)$, если для некоторой перестановки номеров компонент X $\alpha = (\alpha(1), \ldots, \alpha(p))$ для каждого $i = 1, 2, \ldots, p$ найдется $J(i)$ — такое множество из $\{0, 1, \ldots, p\}$, что

$$J(\alpha(i)) = \{\alpha(l_j) : l_j < i \;\; j = 1, \ldots, k_i \leqslant k\}, \tag{4.21}$$

и почти для всех возможных значений $x^{(\alpha(1))}, \ldots, x^{(\alpha(i-1))}$ для всех z

$$P\{x^{(\alpha(i))} < z \mid x^{(\alpha(1))}, \ldots, x^{(\alpha(i-1))}\} = P\{x^{(\alpha(i))} < z \mid X^{(J(\alpha(i)))}\}, \tag{4.22}$$

где $X^{(J(\alpha(i)))} = \{x^{(j)} : j \in J(\alpha(i))\}$ и $x^{(0)} \equiv 1$.

Из этого определения немедленно получаем, что если распределение X непрерывно и удовлетворяет условию $R(k)$, то в обозначениях (4.2′)

$$f(X) = \prod_{1 \le i \le p} f(x^{(i)} \mid X^{(J(i))}). \tag{4.23}$$

При $k = 1$ из (4.21) — (4.23) получаем (4.7), (4.8), т. е. распределение, удовлетворяющее условию $R(1)$, есть распределение с ДСЗ.

Для распределений, удовлетворяющих условию $R(k)$, так же как в § 4.2, можно ввести понятие графа структуры зависимостей G, положив

$$V(G) = \{0, 1, \ldots, p\}, \; E(G) = \left\{\bigcup_i \bigcup_j (i, j) : j \in J(i)\right\}.$$

В отличие от случая $k = 1$ в общем случае $k > 1$ граф структуры зависимостей зависит от выбора перестановки α и не определяется однозначно.

4.4.2. Нормальное $R(k)$-распределение. Пусть I, J — два упорядоченных непересекающихся подмножества $I = \{i_1, \ldots, i_l\}$, $J = \{j_1, \ldots, j_k\}$ координат вектора X. Образуем вектор

$$X^{(I,J)} = (x^{(i_1)}, \ldots, x^{(i_l)}, x^{(j_1)}, \ldots, x^{(j_k)})' = \begin{pmatrix} X^{(I)} \\ X^{(J)} \end{pmatrix},$$

и пусть его корреляционная матрица

$$\Sigma_{(I \cup J)(I \cup J)} = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{JI} & \Sigma_{JJ} \end{bmatrix}.$$

По аналогии с п. 4.2.3 имеем

$$\mathbf{E}(x^{(i)} \mid X^{(J(i))}) = m^{(i)} + \Sigma_{i, J(i)} \Sigma_{J(i) J(i)}^{-1} (X^{(J(i))} - M^{(J(i))});$$

$$\mathbf{D}(x^{(i)} \mid X^{(J(i))}) = \sigma_i^2 - \Sigma_{i, J(i)} \Sigma_{J(i) J(i)}^{-1} \Sigma_{J(i), i} \equiv \sigma_{i, J(i)}^2.$$

Откуда с учетом (4.23) получаем

$$f(X) = (2\pi)^{-p/2} \prod_i \sigma_{i, J(i)}^{-1} \exp\{-[x^{(i)} - m^{(i)} - \Sigma_{i, J(i)} \Sigma_{J(i) J(i)}^{-1} (X^{(J(i))} - M^{(J(i))})]^2 / 2\sigma_{i, J(i)}^2\}. \tag{4.24}$$

Так же, как при $k = 1$, матрица Σ^{-1} имеет очень простой вид. В случае когда α совпадает с исходной нумерацией координат X, в каждом столбце Σ^{-1} над главной диагональю стоит не более k отличных от нуля элементов. Пусть

$$\mathbf{C}(k) = \| c_{ij}(k) \|, \text{ где } c_{ii}(k) = \sigma_{i,J(i)}^{-1};$$

$$\mathbf{C}_{i,J(i)}(k) = -\Sigma_{i,J(i)} \cdot \Sigma_{J(i),J(i)}^{-1} \sigma_{i,J(i)}^{-1},$$

$$c_{il}(k) = 0, \; l \neq \{i\} \cup J(i),$$

тогда имеет место аналог (4.12). Для гауссовских $R(k)$-распределений

$$\Sigma^{-1} = \mathbf{C}'(k)\,\mathbf{C}(k). \tag{4.25}$$

При описании алгоритма выделения графа структуры зависимостей нам потребуется также следующее определение

Определение 4.8. Граф G с $V(G) = \{1, \ldots, p\}$ удовлетворяет условию $T(k)$, если существует хотя бы одна перестановка номеров вершин $\{1, \ldots, p\}$ $\alpha = (\alpha(1), \ldots, \alpha(p))$, что для каждого $1 \leqslant i \leqslant p$ найдется не более k вершин $\alpha(l_1(i)), \ldots, \alpha(l_{k_i}(i))$ $(k_i \leqslant k)$, таких, что для всех $j = 1, \ldots, k_i$; $l_j(i) < i$; $(\alpha(i); \alpha(l_j(i)) \in E(G)$ и $\bigcup\limits_i \bigcup\limits_j (\alpha(i), \alpha(l_j(i)) = E(G)$.

4.4.3. Восстановление графа структуры зависимостей. Пусть $r_{ik \cdot l}$ — *частный коэффициент корреляции между* $x^{(i)}$ и $x^{(k)}$ *при фиксированном* значении $x^{(l)}$ (см. § 1.2 и [20, § 2.5]). Тогда в случае нормального распределения с ДСЗ для $(i, k) \in E(G)$— дереву структуры зависимостей и $(i, j) \notin E(G)$

$$\min_l |r_{ik.l}| \neq 0, \; \min_l |r_{ij.l}| = 0. \tag{4.26}$$

Это свойство после необходимого обобщения может быть использовано для выделения графа структуры зависимостей в случае $R(k)$-распределений. Пусть $r_{ik \cdot J}$ — частный коэффициент корреляции между $x^{(i)}$, $x^{(k)}$ при фиксированном значении $x^{(l_1)}, \ldots, x^{(l_k)}$, $J = \{l_1, \ldots, l_k\}$; назовем *k-весом связи* (i, j)

$$\delta_k(i, j) = \min_J |r_{ij.J}|, \tag{4.27}$$

где минимум берется по всем наборам из k координат X, отличных от $x^{(i)}$ и $x^{(j)}$.

Теорема 4.5. Для невырожденных нормальных $R(k)$-распределений граф структуры зависимостей единствен с точностью до связей нулевого k-веса.

Перейдем к описанию алгоритма выделения графа структуры зависимостей. По своему содержанию он близок к описанному в п. 4.3.2 алгоритму Крускала, только понятие ребра графа, образующего цикл с уже выделенными ребрами, приходится заменить более сложной конструкцией.

Обобщенный алгоритм Крускала. Выбираем на первом шаге ребро l_1 наибольшего k-веса; определяем по индукции последовательность ребер $l_2, l_3, \ldots, l_{n-1}$, выбирая на каждом шаге ребро с наибольшим k-весом, отличное от уже выбранных и обладающее тем свойством, что при добавлении l_n-го к отобранным ребрам граф $(\{1, \ldots, p\}, \{l_1, \ldots, l_n\})$ будет обладать свойством T (k).

В том случае, когда граф структуры зависимостей единствен с точностью до связей нулевого k-веса (нормальные R (k)-распределения), обобщенный алгоритм Крускала дает возможность его восстановить.

4.5. Структура связей нормального вектора *(общий случай)*

С важными, но частными моделями структуры связей между компонентами многомерного нормального вектора мы познакомились в предшествующих параграфах. Наша цель — дать краткую сводку основных результатов общей теории [40, 56, 179].

4.5.1. Марковская тройка. Структура многомерного вектора. Пусть $X = (x^{(1)}, \ldots, x^{(p)})'$ имеет невырожденное p-мерное распределение; $V = \{1, \ldots, p\}$ — множество номеров координат X; A, B, C — непересекающиеся подмножества V; $X^{(A)}$ — подмножество координат X, номера которых входят в A.

Определение 4.9. Тройка (A, B, C) называется марковской, если

$$f(X^{(A)} \mid X^{(B)}, X^{(C)}) = f(X^{(A)} \mid X^{(B)}). \qquad (4.28)$$

В определении марковской тройки допускается тривиальный случай $C = \varnothing$. Для того чтобы тройка (A, B, C) была марковской [61], необходимо и достаточно, чтобы

$$\Sigma_{AC} = \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BC}, \qquad (4.29)$$

где $\Sigma_{AC} = \mathbf{E}(X^{(A)} - \mathbf{E}X^{(A)})(X^{(B)} - \mathbf{E}X^{(B)})'$ или, что эквивалентно,

$$\Sigma^{AC} = 0, \qquad (4.30)$$

где Σ^{AC} (A, C) — блок матрицы Σ^{-1}, соответствующий блоку Σ_{AC} в матрице Σ.

Условие (4.30), очевидно, обобщает соответствующие утверждения о нулях Σ^{-1} в случае R (k)-распределений. В [61] предложен статистический критерий для проверки гипотезы (4.30), построенный в традиционной асимптотике, когда фиксирована матрица Σ, а число наблюдений $n \to \infty$.

Определение 4.10. *Структурой связей* многомерного невырожденного нормального вектора X называется граф $G = (V, E)$, такой, что для любой марковской тройки (i, B, j): а) любая цепь в G из i в j проходит через B и б) для каждого $k \in B$ существует в G цепь из i в j, проходящая через k.

Пусть $\Gamma(i)$ — множество вершин, смежных на G вершине i, т. е. $\Gamma(i) = \{j: (i, j) \in E\}$, тогда

$$f(x^{(i)} \mid X^{(\Gamma(i))}) = f(x^{(i)} \mid X^{(V \setminus i)}), \tag{4.31}$$

причем $\Gamma(i)$ минимально в том смысле, что ни для какого подмножества его компонент (4.31) не имеет места.

Теоретический способ отыскания E состоит в том, что для каждой пары компонент (i, j) подсчитывается частный коэффициент корреляции между i и j при фиксированных значениях всех других компонент [20, § 2.5]. Если он не равен нулю, то $(i, j) \in E$, в противном случае $(i, j) \notin E$. На практике, по-видимому, можно задавать некоторый порог $\delta > 0$ и считать связь $(i, j) \in E$, если $|r_{ij \cdot V \setminus \{i,j\}}| \geqslant \delta$, и $(i, j) \notin E$ — в противном случае. При другом способе все частные коэффициенты корреляции при фиксированных значениях всех других компонент располагают в вариационный ряд по абсолютным величинам и отбирают наперед заданное число наибольших из них. Если (i, j) соответствуют отобранным членам вариационного ряда, то принимают совокупность $(i, j) \in E$. Статистические свойства этих рекомендаций не изучены.

4.5.2. Информационная интерпретация структуры связей. Математическое выражение количества информации в векторе X относительно вектора Y определяется [75] как

$$I(X, Y) = \iint f(X, Y) \ln \frac{f(X, Y)}{f(X) f(Y)} \, dX dY,$$

где $f(X)$, $f(Y)$, $f(X, Y)$ — соответственно плотности распределения X, Y и (X, Y). В нормальном случае информация, заключенная в подмножестве $X^{(A)}$ координат вектора $X \in$

$\in N(0,\Sigma)$ относительно подмножества координат $X^{(B)}$ ($A \cap B = \varnothing$), задается выражением ([75], гл.9, формула (7.4))

$$I\,(X^{(A)},\, X^{(B)}) = \frac{1}{2}\ln \frac{\det \Sigma_{AA} \det \Sigma_{BB}}{\det \Sigma_{(A\cup B)\,(A\cup B)}}. \tag{4.32}$$

Пусть $\Gamma\,(i)$ определено, как в предыдущем пункте, тогда [56]

$$I\,(x^{(i)},\, X^{(\Gamma\,(i))}) > I\,(x^{(i)},\, X^{(B)}), \tag{4.33}$$

где B — любое подмножество компонент X, не содержащее $X^{(\Gamma(i))}$ целиком. Сравним (4.32) с формулой (20) [20, п. 2.5.2]. Из сравнения и (4.33) следует, что максимум взаимной информации между $x^{(i)}$ и $X^{(A)}$ достигается на том же наборе компонент $X^{(\Gamma(i))}$, что и максимум коэффициента множественной корреляции. При этом структура связей выделяет не парные, а множественные зависимости, в большей степени отражающие реальное взаимодействие переменных. Для каждой компоненты $x^{(i)}$ с помощью графа структуры связей легко находится группа координат $X^{(\Gamma(i))}$, непосредственно связанных с $x^{(i)}$ и несущих максимальную информацию о ней.

4.5.3. Использование структуры для представления распределения в виде композиции более простых распределений. Начнем с формулы (4.4). Разобьем каждое из подмножеств компонент, входящих в правые части сомножителей (4.4), на два таких подмножества $\{X^{(A_1)},\ X^{(B_1)}\}$, $\{X^{(A_2)},\ X^{(B_2)}\}, \ldots,$ $\{X^{(A_{p-1})}, X^{(B_{p-1})}\}$, что тройки $(1,\, A_1,\, B_1), \ldots, (p,\, A_{p-1},\, B_{p-1})$ будут марковскими. В результате получаем аналог разложений (4.5) и (4.23):

$$f\,(X) = f\,(x^{(1)})\, f\,(x^{(2)}\mid X^{(A_1)}) \ldots f\,(x^{(p)}\mid X^{(A_{p-1})}).$$

К сожалению, в общем случае так же, как для $R\,(k)$-распределений, не выработано простых рекомендаций, как наиболее удачным образом с точки зрения простоты окончательной формулы выбрать первоначальный порядок координат.

ВЫВОДЫ

1. При содержательной интерпретации взаимозависимостей между координатами случайного вектора целесообразно выделять *связи прямые* и *опосредованные*. Важным примером непосредственной связи является связь последовательных наблюдений $(\xi_n,\ \xi_{n+1})$ в цепи Маркова. Связь наблюдений $(\xi_n,$ $\xi_{n+k})$ опосредуется через наблюдения $(\xi_{n+1},\ \xi_{n+2},\ \ldots,\ \xi_{n+k-1})$.

Для визуального представления зависимостей широко используются *графы структуры зависимостей*, в которых координаты вектора изображаются в виде вершин графа, а непосредственные связи между ними — в виде связывающих их ребер.

2. Понятие *древообразной структуры зависимостей* между координатами случайного вектора возникает как обобщение понятия марковости для совокупности случайных величин, лишенных временной упорядоченности. Говорят, что распределение $X = (x^{(1)}, \ldots, x^{(p)})'$ имеет ДСЗ, если существует такая перестановка координат вектора $(\alpha(1), \ldots, \alpha(p))$, что

$$f(X) = \prod_i f(x^{\alpha(i)} \mid x^{j(\alpha(i))}),$$

где $j(\alpha(i)) \in \{\alpha(1), \ldots, \alpha(i-1)\}$.

3. Невырожденные p-мерные нормальные распределения с ДСЗ имеют очень простой вид матрицы Σ^{-1}, где Σ — ковариационная матрица координат вектора. В Σ^{-1} над главной диагональю стоит не более $p - 1$ отличных от нуля элементов. Эта малопараметричность описания ковариационной матрицы в сочетании с большим разнообразием описываемых классов зависимостей, включающим, в частности, все ковариационные матрицы цепей Маркова, делает распределения с ДСЗ одним из основных инструментов в многомерном анализе.

4. Для распределений с ДСЗ при выполнении дополнительного условия, справедливого для всех невырожденных нормальных распределений, графы структуры зависимостей определяются однозначно с точностью до связей, соответствующих независимым координатам. С другой стороны, для этих распределений по графу структуры зависимостей восстанавливается, хотя и неоднозначно, естественный порядок координат, фигурирующий в определении распределений с ДСЗ.

5. Если известна корреляционная матрица невырожденного нормального вектора с ДСЗ, то по ней с помощью известного в теории графов алгоритма Крускала граф структуры зависимостей восстанавливается однозначно. Алгоритм Крускала, примененный к выборочной корреляционной матрице, оказывается состоятельным в асимптотике Колмогорова — Деева, специально рассчитанной на изучение ситуаций, когда число наблюдений вектора и его размерность суть величины одного порядка.

6. $R(k)$-распределения ($k \geqslant 1$) возникают как результат обобщения, с одной стороны, распределений с ДСЗ ($R(1)$-распределений), а с другой — k-зависимых марковских последовательностей. На $R(k)$-распределения удается перенести многие свойства распределений с ДСЗ.

7. Пусть A, B, C — непересекающиеся подмножества номеров координат, а $X^{(A)}$, $X^{(B)}$, $X^{(C)}$ — соответствующие наборы координат. Тройка (A, B, C) называется марковской, если $f(X^{(A)} \mid (X^{(B)}, X^{(C)}) = f(X^{(A)} \mid X^{(B)})$. Построены статистические критерии для проверки гипотезы, что заданная тройка — марковская. В случае когда X имеет невырожденное нормальное распределение, структурой связей X называется граф G, вершинами которого являются номера координат X, а ребрами — соединяющие их дуги и для которого выполняется условие, что для каждой марковской тройки (i, B, j): а) любая цепь в G из i в j проходит через B и б) для каждого $k \in B$ существует цепь в G из i в j, проходящая через k. Вся информация в координатах $X \setminus x^{(i)}$ относительно координаты $x^{(i)}$ содержится только в $X^{(\Gamma(i))}$, где $\Gamma(i)$ — вершины графа G, смежные с вершиной i. Ребрам (i, j) графа G соответствуют отличные от нуля частные коэффициенты корреляции между i и j при фиксированных остальных координатах вектора X. Этот факт можно использовать для нахождения графа структуры связей.

Раздел II. ИССЛЕДОВАНИЕ ВИДА ЗАВИСИМОСТИ МЕЖДУ КОЛИЧЕСТВЕННЫМИ ПЕРЕМЕННЫМИ

(регрессионный анализ)

Глава 5. ОСНОВНЫЕ ПОНЯТИЯ РЕГРЕССИОННОГО АНАЛИЗА

Предыдущий раздел (гл. 1—4) посвящен описанию математического аппарата, привлекаемого для реализации 3-го этапа статистического исследования зависимостей (см. «Корреляционный анализ» в п. В.6), на котором исследователь пытается проанализировать *структуру связей* между рассматриваемыми переменными и *измерить степень их тесноты*. После того как он убедится в наличии статистически значимых связей между анализируемыми переменными, он приступает к выявлению и математическому описанию *конкретного вида* интересующих его зависимостей: подбирает класс функций, в рамках которого будет вести свой дальнейший анализ (этап 4); производит, если это необходимо, отбор наиболее информативных предсказывающих переменных (этап 5); вычисляет оценки для неизвестных значений параметров, участвующих в записи уравнения искомой зависимости (этап 6); анализирует точность полученного уравнения связи (этап 7). Этапы 4—7 и составляют содержание *регрессионного анализа*, описанию которого посвящен данный раздел.

Но прежде чем переходить к изложению методов, составляющих аппарат регрессионного анализа, необходимо ввести и прокомментировать ряд основных понятий и определений.

5.1. Функция регрессии как условное среднее и ее интерпретация в рамках многомерной нормальной модели

Во введении при общей формулировке задачи статистического исследования зависимостей (п.В.1), при описании основных прикладных проблем, в решении которых используется аппарат статистического исследования зависимостей (п. В.4), и при классификации основных типов исследуемых зависимостей (п. В.5) мы, по существу, уже использовали понятие «функции

регрессии». Перед тем как сформулировать общее определение функции регрессии, вернемся к примерам В.1 и В.2

В примере В.1 мы исследовали, как меняется *средняя* величина удельных денежных сбережений семьи (η) в зависимости от ее среднедушевого дохода (ξ), причем усреднение денежных сбережений (η) производилось по всем семьям данной группы по доходам (т. е. при $\xi = x$). Другими словами, анализировалась зависимость *условного среднего значения* удельных семейных сбережений $y_{\text{ср}}(x) = \mathbf{E}(\eta \mid \xi = x)$ от среднедушевого дохода x (см. табл. В.1 и рис. В.2).

В примере В.2 анализировалось поведение показателя средней долговечности (η) испытуемого образца в зависимости от величины характеристики эксплуатационного напряжения (x), где усреднение величины η производилось по всем образцам, испытанным при заданном значении характеристики эксплуатационного напряжения x. Таким образом, речь опять идет об исследовании зависимости условного среднего значения результирующего показателя η (вычисленного при условии, что объясняющая переменная приняла заданное значение x) от текущего значения объясняющей переменной (см. табл. В.4 и рис. В.5).

Рассмотрим общую схему. Пусть значение исследуемого результирующего показателя η при данных *фиксированных* величинах объясняющих переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ случайным образом флюктуирует вокруг некоторого (вообще говоря, неизвестного) уровня $f(x^{(1)}, x^{(2)}, \ldots, x^{(p)})$, зависящего от конкретных значений предикторов $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, т. е.

$$\eta = f(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) + \varepsilon(x^{(1)}, \ldots, x^{(p)}), \qquad (5.1)$$

где остаточная компонента $\varepsilon(X)$ определяет случайное отклонение значения η от постоянного (при фиксированных $x^{(1)}, \ldots, x^{(p)}$) уровня f. При этом наличие флюктуации ε может быть присуще *самой природе эксперимента или наблюдения* (как в примерах В.1 и В.2), а может объясняться случайными ошибками в измерении величины f (тогда η является результатом несколько искаженного измерения значения f). Когда говорят, что «некоторая величина (η) случайным образом флюктуирует вокруг определенного (неслучайного) уровня f», то, как правило, имеют в виду, что среднее значение такой флюктуирующей случайной величины должно быть равно f, т. е. $\mathbf{E}\eta = f$. Поскольку условия эксперимента и, в частности, уровень, около которого флюктуирует η, зависят от *конкретных* значений $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ некоторого набора объясняющих переменных, со-

ответственно $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}$, то из (5.1) и только что сказанного непосредственно следует

$$\mathbf{E}(\eta \mid \xi^{(1)} = x^{(1)}, \ldots, \xi^{(p)} = x^{(p)}) = f(x^{(1)}, \ldots, x^{(p)}). \quad (5.2)$$

Функция $f(x^{(1)}, x^{(2)}, \ldots, x^{(p)})$, описывающая зависимость условного среднего значения $y_{ср}(X)$ результирующего показателя η (вычисленного при условии, что величины предсказывающих переменных зафиксированы на уровнях $x^{(1)}, x^{(2)}, \ldots, \ldots, x^{(p)}$) от заданных фиксированных значений предсказывающих переменных, называется функцией регрессии.

В общем случае для точного описания функции регрессии необходимо точное знание условного закона распределения результирующего показателя η (при условии, что $\xi = X$). Поскольку в статистической практике мы никогда не располагаем такой информацией, то обычно ограничиваются поиском *подходящих аппроксимаций* для $f(X)$, основанных на исходных статистических данных вида (В.1) (о методах построения таких аппроксимаций см. гл. 7—10).

Однако в жестких теоретических рамках модельных допущений о типе распределения исследуемого вектора показателей $(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}; \eta)$ может быть получен общий вид функции регрессии $f(X) = \mathbf{E}(\eta \mid \xi = X)$ (здесь, как и ранее, $\xi = (\xi^{(1)}, \ldots, \xi^{(p)})'$ и $X = (x^{(1)}, \ldots, x^{(p)})'$). Так, например, если предположить, что исследуемый вектор переменных $(\xi'; \eta)'$ подчиняется $(p+1)$-мерному нормальному распределению с вектором средних значений

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_\xi \\ m^{(\eta)} \end{pmatrix}, \text{ где } \mathbf{M}_\xi = \begin{pmatrix} m^{(1)} \\ \vdots \\ m^{(p)} \end{pmatrix},$$

и с ковариационной матрицей

$$\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\eta} \\ \Sigma'_{\xi\eta} & \sigma_{\eta\eta} \end{pmatrix},$$

где

$$\Sigma_{\xi\xi} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}, \quad \Sigma_{\xi\eta} = \begin{pmatrix} \sigma_{1\eta} \\ \sigma_{2\eta} \\ \vdots \\ \sigma_{p\eta} \end{pmatrix},$$

а

$$\sigma_{ij} = \mathbf{E}(\xi^{(i)} - m^{(i)})(\xi^{(j)} - m^{(j)}),$$

$$\sigma_{i\eta} = \mathbf{E}(\xi^{(i)} - m^{(i)})(\eta - m^{(\eta)}), \quad \sigma_{\eta\eta} = \mathbf{E}(\eta - m^{(\eta)})^2,$$

то из (1.3) непосредственно следует

$$f(X) = \mathbf{E}(\eta \mid \xi = X) = m^{(\eta)} + \Sigma'_{\xi\eta} \cdot \Sigma^{-1}_{\xi\xi} \cdot (X - M_\xi). \qquad (5.3)$$

Таким образом, если анализируемый многомерный признак $(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}; \eta)$ подчинен $(p+1)$-мерному нормальному закону, то функция регрессии результирующего показателя η по объясняющим переменным $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}$ имеет *линейный* (*по* X) *вид*, а ее коэффициенты выражаются в терминах первых двух моментов анализируемых случайных величин.

Происхождение термина «регрессия» (лат. «regression» — отступление, возврат к чему-либо) связано только с прикладной спецификой одного из первых конкретных примеров, в котором это понятие было использовано, но никак не с его общесмысловым наполнением. Этот термин был введен английским психологом и антропологом Ф. Гальтоном в связи с вопросом о наследственности роста. Обрабатывая статистические данные, Гальтон нашел, что сыновья отцов, отклоняющихся по росту на x дюймов от среднего роста всех отцов, сами отклоняются от среднего роста всех сыновей меньше, *чем на* x *дюймов*. Гальтон назвал выявленную тенденцию «регрессией к среднему состоянию» («regression to mediocrity»). Однако термин столь прочно внедрился в статистическую литературу, что мы не делаем попытки заменить его более подходящим для выражения существенных свойств понятия статистической зависимости.

5.2. Функция Δ-регрессии как решение оптимизационной задачи

В предыдущем параграфе обращается внимание читателя на то, что в статистической практике приходится ограничиваться поиском *подходящих аппроксимаций* для неизвестной истинной функции регрессии $f(X)$, поскольку исследователь не располагает точным знанием условного закона распределения вероятностей анализируемого результирующего показателя η (при условии, что объясняющие переменные ξ приняли «значение», равное X).

В данном параграфе будет уточнено, что значит «подходящая аппроксимация», т. е. будут описаны *критерии адекватности модели*, в соответствии с которыми естественно измерять качество предполагаемой аппроксимации $f_a(X)$ искомой функции регрессии $f(X)$ в том или ином случае.

Общий оптимизационный подход к построению статисти-

ческих решающих процедур описан в [13] и кратко воспроизведен в [14, § 1.2].

Остановимся на конкретизации этого подхода применительно к задачам статистического исследования зависимостей и, в частности, к задаче наилучшего восстановления (по исходным статистическим данным вида (B.1)) условного значения результирующего показателя $\eta(X) = (\eta \mid \xi = X)$ и неизвестной функции регрессии $f(X) = \mathbf{E}(\eta \mid \xi = X)$. С этой целью воспользуемся следующей схемой рассуждений.

а. Введем *функцию потерь* $\rho(\widehat{\varepsilon_{f_a}}(X))$, измеряющую убытки от неточности восстановления значения $\eta(X) = \{\eta \mid \xi = X\}$ с помощью функции $f_a(X)$; здесь $\widehat{\varepsilon_{f_a}} = \eta(X) - f_a(X)$, а функция $\rho(u)$, как правило, монотонно неубывающая, чаще всего выпуклая, функция аргумента u с неотрицательными значениями (см. различные варианты функции ρ в § 7.2).

б. Определим *теоретический* и соответствующий ему *выборочный критерии адекватности модели* $f_a(X)$, используемой в качестве аппроксимации для неизвестного условного значения результирующего показателя $\eta(X) = (\eta \mid \xi = X)$:

$$\text{теоретический } \Delta(f_a) = \mathbf{E}\rho(\widehat{\varepsilon_{f_a}}(X)); \tag{5.4}$$

$$\text{выборочный } \widehat{\Delta}_n(f_a) = \frac{1}{n}\sum_{i=1}^{n}\rho(\widehat{\varepsilon_{f_a}}(X_i)). \tag{5.4'}$$

В (5.4) усреднение производится и по всем возможным значениям случайной величины $\widehat{\varepsilon_{f_a}}(X)$ (при каждом фиксированном X) и по всем возможным значениям X, а в (5.4′) — по всем имеющимся наблюдениям.

в. Зададимся *классом допустимых решений* $\mathbf{F}$, в рамках которого будем вести дальнейший поиск наилучшей, в смысле критериев Δ или $\widehat{\Delta}_n$, аппроксимации f_a^* (или $\widehat{f_a^*}$) для $\eta(X)$. При этом если в качестве класса $\mathbf{F}$ задаются некоторым *параметрическим семейством функций*

$$\mathbf{F}_\Theta = \{f_a(X;\ \Theta)\}_{\Theta \in \Gamma}, \tag{5.5}$$

то задача подбора наилучшей аппроксимации f_a^* (или $\widehat{f_a^*}$) сводится к определению таких значений параметров Θ^* (или $\widehat{\Theta}^*$), при которых некоторая агрегированная характеристика точности восстановления значений $\eta(X)$ по значениям $f_a(X;\ \Theta)$ (или $\widehat{f_a}(X;\ \Theta)$) является наилучшей (подход, основанный на ис-

пользовании в качестве класса допустимых решений F параметрических семейств вида (5.5) называют *параметрическим*).

г. Будем называть функцию $f_a^\Delta(X)$ *функцией* Δ-*регрессии, если она дает прогноз для условных значений результирующего показателя* $\eta(X)$, *являющийся наилучшим в смысле критерия адекватности* Δ. Другими словами:

$$f_a^\Delta(X)=\arg\min_{f_a\in \mathbf{F}}\Delta(f_a). \qquad (5.6)$$

Покажем (на примере квадратичной функции потерь, т. е. при $\rho(u)=u^2$), что задача минимизации функционала (5.4) содержит задачу наиболее точного восстановления регрессии. Действительно, для критерия (5.4) справедливо тождество (см. п. 1.3.1)

$$\Delta(f_a)=\int\limits_X\int\limits_y (y-f_a(X))^2 p_\eta(y\mid X)\,p_\xi(X)\,\mathrm{d}y\mathrm{d}X=$$

$$=\int\limits_X\int\limits_y (y-f(X))^2 p_\eta(y\mid X)\,p_\xi(X)\,\mathrm{d}y\mathrm{d}X+$$

$$+\int\limits_X (f_a(X)-f(X))^2 p_\xi(X)\,\mathrm{d}X$$

(здесь $p_\eta(y|X)$ и $p_\xi(X)$ — соответственно условная функция плотности результирующего показателя η при условии, что $\xi=X$, и частная функция плотности предикторной переменной ξ).

Так как первое слагаемое в правой части этого тождества *не зависит* от функции $f_a(X)$, то минимум функционала $\Delta(f_a)$ определяется величиной второго слагаемого и достигается на такой функции $f_a^\Delta(X)\in \mathbf{F}$, на которой минимизируется погрешность описания истинной функции регрессии $f(X)$ с помощью функций из класса F.

В дальнейшем, чтобы отличать *теоретическую* версию этого определения (которая соответствует функционалу (5.4)) от *выборочной* (функционал (5.4′)) и с целью упрощения обозначений, будем полагать (если не требуется специальных пояснений, связанных с выбором критерия Δ)

$$f_a(X)=f_a^\Delta(X); \qquad (5.7)$$

$$\widehat{f}_a(X)=f_a^{\widehat{\Delta_n}}(X) \qquad (5.7')$$

и называть их соответственно *теоретической* и *выборочной аппроксимациями* истинной функции регрессии. Основанием для

подобной терминологии служат простые асимптотические соотношения, связывающие в ряде достаточно общих случаев функции $f(X) = \mathbf{E}(\eta \mid \xi = X)$, $f_a(X)$ и их выборочные аналоги (см. следующий параграф).

Обратим внимание читателя на ряд частных случаев функции потерь $\rho(u)$, широко используемых в теории и практике статистического исследования зависимостей:

1) $\rho(u) = u^2$; получаемая в соответствии с (5.6) регрессия называется *среднеквадратической*, а метод, реализующий минимизацию функционала $\widehat{\Delta}_n(f_a)$, принято называть *методом наименьших квадратов* (см. § 7.1);

2) $\rho(u) = |u|$; получаемая в соответствии с (5.6) регрессия называется *среднеабсолютной* (или *медианной*), а метод, реализующий минимизацию функционала $\widehat{\Delta}_n(f_a)$, называют *методом наименьших модулей* (см. п.7.2.1);

3) $\rho(u) = |u|^\tau$, где $\tau \to \infty$; можно показать, что в этом случае минимизация критерия $\widehat{\Delta}_n(f_a)$ сводится к минимизации (по $f_a \in \mathbf{F}$) $\max\limits_{1 \leqslant i \leqslant n} |y_i - f_a(X_i)|$, поэтому соответствующую регрессию называют *минимаксной*.

Другие важные частные случаи Δ-регрессии читатель найдет в § 7.2.

5.3. Взаимоотношения различных регрессий

Взаимоотношения истинной и Δ-регрессий существенно зависят от вероятностной природы регрессионных остатков $\varepsilon(X)$ в моделях типа (5.1) и от способа выбора класса допустимых решений $\mathbf{F}$. Попробуем вначале понять эти взаимоотношения на примере.

Пример 5.1. Результирующий показатель η связан с объясняющей переменной ξ соотношением

$$\eta = 2 \cdot \xi^{1,5} + \varepsilon, \tag{5.8}$$

где регрессионный остаток ε — случайная величина, подчиняющаяся нормальному закону распределения со средним значением $\mathbf{E}\varepsilon = 0$ и с дисперсией $\mathbf{D}\varepsilon = 4$, а диапазон возможных значений ξ определяется отрезком [2; 10]. Очевидно, истинная функция регрессии в данном случае имеет вид

$$f(x) = \mathbf{E}(\eta \mid \xi = x) = 2x^{1,5}. \tag{5.9}$$

Предположим, нам *не известен точный вид соотношения* (5.8) и соответственно не известно уравнение функции регрес-

сии (5.9). Однако мы располагаем следующей системой двумерных наблюдений $(x_i, \ y_i)_{i=\overline{1,9}}$, генерируемых моделью (5.8), т. е. связанных соотношением $y_i = 2x_i^{1,5} + \varepsilon_i$ (табл. 5.1)

Таблица 5.1

Номер наблюдения (i)	1	2	3	4	5	6	7	8	9
(x_i) i-е наблюденное значение ξ	2	3	4	5	6	7	8	9	10
(y_i) i-е наблюденное значение η	6,58	10,67	20,91	21,71	29,25	37,64	44,67	56,60	61,33

Расположение точек — наблюдений на рис. 5.1 дает нам основание ограничить класс допустимых решений только линейными зависимостями, т. е. определить в качестве класса допустимых решений параметрическое семейство

$$\mathbf{F}_{\text{лин}} = \{\theta_0 + \theta_1 x\}. \qquad (5.10)$$

Имея априорную информацию о типе распределения регрессионных остатков, остановим свой выбор на *квадратичной* функции потерь. Решая оптимизационную задачу вида

$$\sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \to \min_{\theta_0, \theta_1}, \qquad (5.11)$$

получаем оценки $\widehat{\theta_0^*}$, $\widehat{\theta_1^*}$ для неизвестных параметров θ_0, θ_1, участвующих в записи аппроксимирующей функции $f_a(x) = \theta_0 + \theta_1 x$. График соответствующей *выборочной аппроксимирующей функции регрессии* $\widehat{f_a}(x) = \widehat{\theta_0^*} + \widehat{\theta_1^*} \cdot x$ изображен на рис. 5.1. Для сравнения на том же рисунке изображены графики *истинной функции регрессии* $f(x) = 2x^{1,5}$ и *теоретической аппроксимирующей функции регрессии* $f_a(x)$. Последняя характеризует результат, к которому мы бы неограниченно приближались (в смысле сходимости по вероятности), решая оптимизационную задачу (5.11) для неограниченно расширяющейся по объему выборки $\{(x_i, y_i)\}_{i=\overline{1,n}}$, $n \to \infty$. Поскольку мы «не угадали» класс допустимых решений (истинная функция регрессии не принадлежит к выбранному нами классу (5.10)), то в данном случае мы находимся в ситуации (к сожалению, достаточно типичной для практики статистических исследова-

ний), в которой наши статистические *выводы и оценки не будут обладать свойством состоятельности*. Другими словами, как бы мы ни увеличивали объем исходной статистической базы, мы не сможем добиться сходимости нашей выборочной аппроксимирующей функции регрессии $\widehat{f}_a(x)$ к истинной функции регрессии $f(x)$.

Напротив, если бы мы правильно выбрали класс допустимых решений, что в данном примере означало бы

$$\mathbf{F}_{\text{ст}} = \{\theta_0 x^{\theta_1}\}, \qquad (5.12)$$

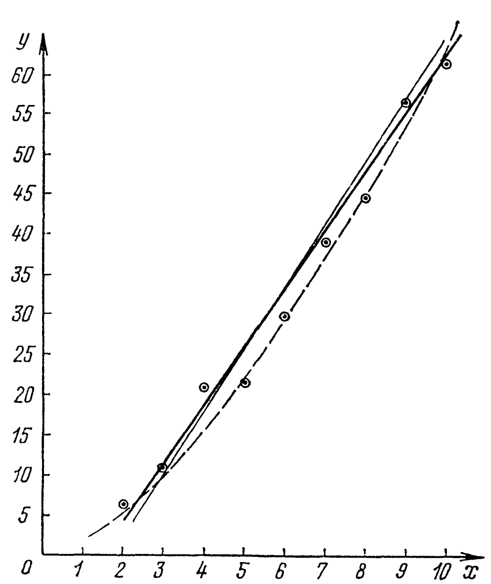


Рис. 5.1. Взаимное расположение истинной, теоретической аппроксимирующей и выборочной аппроксимирующей функций регрессии в примере 5.1:

– – – – – $f(x) = \mathrm{E}(\eta/\xi = x) = 2\,x^{1,5}$;

▬▬▬ – $f_a(x) = \theta_0^* + \theta_1^* x$;

——— – $\widehat{f}_a(x) = \widehat{\theta}_0^* + \widehat{\theta}_1^* x$

то неточность в описании $f(x)$ с помощью $\widehat{f}_a(x)$ объяснялась бы *только ограниченностью выборки*, по которой строится функция $f_a(x)$, и, следовательно, могла бы быть сделана сколь угодно малой за счет $n \to \infty$.

Сформулируем в заключение несколько общих положений, относящихся к сравнению различных функций регрессии:

а) истинная регрессия $f(X) = \mathbf{E}(\eta \mid \xi = X)$ является одновременно среднеквадратической, т. е. дает решение оптимизационной задачи вида (5.6) при квадратичной функции потерь ($\rho(u) = u^2$) и при отсутствии ограничений на класс допустимых решений $\mathbf{F}$ (доказательство приведено в п.1.3.1);

б) для широкого класса критериев адекватности $\Delta(f_a)$ выборочная аппроксимирующая функция регрессии $\widehat{f_a}(X)$ сходится (по вероятности) к теоретической аппроксимирующей функции регрессии $f_a(X)$ при $n \to \infty$;

в) в случае удачного выбора класса допустимых решений, т. е. при $f(X) \in \mathbf{F}$, теоретическая аппроксимирующая функция регрессии $f_a(X)$ (при надлежащем выборе критериев адекватности Δ) совпадает с истинной и соответственно выборочная аппроксимирующая функция регрессии $\widehat{f_a}(X)$ будет сходиться (по вероятности) к истинной;

г) в случае неудачного выбора класса допустимых решений, т. е. при $f(X) \notin \mathbf{F}$, ошибку в описании истинной функции регрессии $f(X)$ с помощью выборочной аппроксимирующей функции регрессии $\widehat{f_a}(X)$ удобно представить в виде суммы двух компонент: *ошибки выборки* и *ошибки аппроксимации*. При этом ошибка выборки (разность $\widehat{f_a}(X) - f_a(X)$) при $n \to \infty$ стремится (по вероятности) к нулю, в то время как ошибка аппроксимации (разность $f_a(X) - f(X)$) *не стремится к нулю* ни при каком выборе критерия адекватности Δ.

Обсуждение мотивов выбора вида функции потерь ρ (и соответственно критерия адекватности Δ) приводится в гл. 7.

ВЫВОДЫ

1. Центральное место в аппарате статистического исследования зависимостей между количественными переменными занимает *понятие регрессии* результирующего показателя η по объясняющим переменным $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}$.
2. Функция $f(X)$, описывающая изменение условного среднего значения $y_{ср}(X) = \mathbf{E}(\eta \mid \xi = X)$ результирующего показателя η в зависимости от изменения заданного значения X предикторной переменной ξ, называется функцией регрессии.
3. Для точного описания функции регрессии $f(X) = \mathbf{E}(\eta \mid \xi = X)$ необходимо знание закона условного распределения результирующего показателя η (при условии $\xi = X$). В статистической практике ограничиваются оценкой (на основании имеющихся выборочных данных вида (В.1)) *подходящих аппроксимаций* $\widehat{f_a}(X)$ функции $f(X)$.

4. Наряду с приведенным выше *классическим определением* функции регрессии в теории и практике статистического исследования зависимостей используются *функции Δ-регрессии*, являющиеся наилучшими прогностическими моделями для анализируемого результирующего показателя $\eta(X)$ в смысле минимизации заданного критерия адекватности (агрегированной ошибки прогноза) $\Delta(f_a)$. Функции Δ-регрессии позволяют подбирать наилучшие аппроксимации для неизвестной истинной функции регрессии. Кроме того, они представляют и самостоятельный интерес, позволяя строить и анализировать иную, чем условное среднее, условную характеристику места группирования результирующего показателя $\eta(X) = (\eta|\xi = X)$, обладающую в ряде ситуаций определенными преимуществами перед условной средней.

5. Наиболее распространенными частными случаями Δ-регрессий являются *среднеквадратическая*, *медианная* и *минимаксная* регрессии. Весьма полезными являются и различные варианты так называемых «робастных» регрессий (см. § 7.2).

6. Соотношение истинной $(f(X))$, теоретической аппроксимирующей $(f_a(X))$ и выборочной аппроксимирующей $(\widehat{f}_a(X))$ регрессий существенно зависит от выбора критерия адекватности $\Delta(f_a)$ (определяемого природой регрессионных остатков ε) и класса допустимых решений **F**. В частности, даже при удачном выборе критерия адекватности Δ в ситуациях, когда истинная функция регрессии $f(X)$ не «накрывается» классом допустимых решений **F** (т. е. когда $f(X) \notin \mathbf{F}$), выборочная аппроксимирующая функция регрессии $\widehat{f}_a(X)$ не будет стремиться к истинной при неограниченном росте объема выборки (отсутствие свойства состоятельности у $\widehat{f}_a(X)$, объясняемое неустранимостью ошибки аппроксимации).

7. *Истинная регрессия* $f(X) = \mathbf{E}(\eta|\xi = X)$ является одновременно *среднеквадратической*, т. е. дает решение оптимизационной задачи вида (5.6) при квадратичной функции потерь (при отсутствии ограничений на класс допустимых решений **F**).

Глава 6. ВЫБОР ОБЩЕГО ВИДА ФУНКЦИИ РЕГРЕССИИ

Собственно регрессионный анализ, т. е. восстановление по имеющимся наблюдениям предикторной переменной ξ и результирующего показателя η

$$\{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\} \quad (6.1)$$

неизвестной функции регрессии $f(X) = \mathbf{E}(\eta | \xi = X)$, начинается с выбора класса допустимых решений $\mathbf{F}$ — класса функций, в рамках которого предполагается вести поиск наиболее подходящей аппроксимации $\widehat{f}_a(X)$ для $f(X)$.

Наиболее распространенными в статистической практике являются *параметрические регрессионные схемы*, когда в качестве класса допустимых решений выбирается некоторое параметрическое семейство функций

$$\mathbf{F} = \{f(X;\ \Theta)\}_{\Theta \in \Gamma}. \tag{6.2}$$

В этом случае дальнейший поиск аппроксимации $\widehat{f}(X)$ сводится к наилучшему (в смысле заданного выборочного критерия адекватности, см. § 5.2) подбору неизвестного значения параметра $\widehat{\Theta}^*$, что в свою очередь осуществляется с помощью *полностью формализованного* алгоритма решения соответствующей оптимизационной задачи, составляющей математическую основу процедуры, называемой *статистическим оцениванием параметра*.

Но до перехода к процедуре статистического оценивания неизвестного значения параметра мы должны сделать и обосновать определенный выбор типа параметрического семейства (6.2). Так, например, в качестве класса допустимых решений можно использовать

линейные функции: $$f(X;\ \Theta) = \theta_0 + \sum_{k=1}^{p} \theta_k \cdot x^{(k)}; \tag{6.2'}$$

степенные функции: $$f(X;\ \Theta) = \theta_0 (x^{(1)})^{\theta_1} (x^{(2)})^{\theta_2} \ldots (x^{p})^{\theta_p}; \tag{6.2''}$$

алгебраические полиномы степени $m \geqslant 2$:

$$f(X;\ \Theta) = \theta_0 + \sum_{k=1}^{p} \theta_k \cdot x^{(k)} + \sum_{k_1=1}^{p} \sum_{k_2=1}^{p} \theta_{k_1 k_2} x^{(k_1)} \cdot x^{(k_2)} + \ldots$$

$$\ldots + \sum_{k_1=1}^{p} \cdots \sum_{k_m=1}^{p} \theta_{k_1 k_2 \ldots k_m} \cdot x^{(k_1)} \cdot x^{(k_2)} \ldots x^{(k_m)} \tag{6.2'''}$$

и т. д.

Следует подчеркнуть, что этап 4 (см. § В.6), т. е. этап исследования, посвященный выбору общего вида функции регрессии (параметризация модели), бесспорно, является *ключевым*: от того, насколько удачно он будет реализован, решающим образом зависит точность восстановления неизвестной функции регрессии $f(X)$. В то же время приходится признать, что этот этап находится, пожалуй, в самом невыгодном положении: к сожалению, не существует системы стандартных рекомендаций

и методов, которые образовывали бы строгую теоретическую базу для его наиболее эффективной реализации.

Остановимся на некоторых рекомендациях, связанных с реализацией трех основных моментов, учет которых необходим при решении проблемы выбора общего вида функции регрессии: 1) максимальное использование априорной информации о *содержательной* (физической, экономической, социологической и т. п.) *сущности* анализируемой зависимости; 2) предварительный анализ *геометрической структуры* исходных данных вида (6.1), на основании которых конструируется искомая зависимость; 3) различные *статистические приемы* обработки исходных данных, позволяющие сделать наилучший выбор из нескольких сравниваемых вариантов.

6.1. Использование априорной информации о содержательной сущности анализируемой зависимости

Анализируя содержательную сущность изучаемой зависимости, исследователь еще *до обращения* к исходным статистическим данным может (и должен!) попытаться ответить на ряд вопросов по поводу характера искомой регрессионной связи:

а) будет ли искомая функция $f(X)$ монотонной или она должна иметь один (или несколько) экстремум?

б) следует ли ожидать стремления (в процессе $x^{(k)} \to \infty$) $f(X)$ к асимптотам (по одной или нескольким предикторным переменным) и какова их содержательная интерпретация? Так, например, если $f(X)$ — средний объем благ определенного вида, потребляемых семьями группы X по доходам, то, очевидно, при $X \to \infty$ следует ожидать «насыщения», т. е. $f(X)$ будет стремиться (снизу) к горизонтальной асимптоте (см. п. 4 и 10 в табл. В.3);

в) какова принципиальная природа воздействия предикторных переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ на формирование результирующего показателя y — аддитивная или мультипликативная? Так, например, многие схемы зависимостей в экономике и квалиметрии характеризуются мультипликативной природой воздействия предикторов на y (см. п. 1—3 в табл. В.3, а также [5]);

г) не диктует ли содержательный смысл анализируемой зависимости обязательное прохождение графика искомой функции $f(X)$ через одну или несколько априори заданных точек в исследуемом факторном пространстве (X, y)?

Поясним необходимость и возможность максимального извлечения информации об общем виде анализируемой функции

регрессии $f(X)$ из соображений профессионально-теоретического характера на двух примерах.

Пример 6.1. На рис. 6.1 представлены 63 результата специального эксперимента [50, с. 57]. Расположение точек на рис. 6.1 не дает ответа на вопрос, описывать ли зависимость между скоростью автомобиля (x миль/ч) и расстоянием (y футов), пройденным им после поданного сигнала об остановке, линейной или параболической зависимостью.

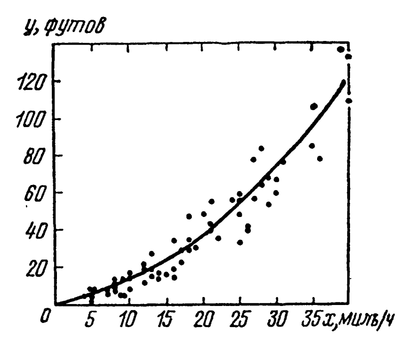


Рис. 6.1. График зависимости тормозного пути автомобиля (y) от скорости его движения (x)

Этот вопрос остается без ответа и после построения соответствующих кривых и применения известных статистических критериев, предназначенных решать, насколько хорошо согласуются кривые с экспериментальными данными. Однако несложные рассуждения профессионально-теоретического характера все-таки позволяют сделать этот выбор. Действительно, для каждого отдельного автомобиля и водителя расстояние, пройденное до остановки, определяется в основном тремя факторами: скоростью автомобиля (x) в момент подачи сигнала об остановке, временем реакции на этот сигнал водителя (θ_1, ч) и тормозами автомобиля. Автомобиль успеет пройти путь $\theta_1 x$ до момента включения водителем тормозов и еще $\theta_2 \cdot x^2$ после этого момента, поскольку согласно элементарным физическим законам теоретическое расстояние, пройденное до остановки с момента торможения, пропорционально квадрату скорости.

Итак, $y = \theta_1 x + \theta_2 x^2$, что после оценивания θ_1 и θ_2 с помощью мнк (см. гл. 7) дает $y = 0{,}76x + 0{,}056x^2$.

Пример 6.2[1]. Рассмотрим в качестве результирующего показателя η вес коровы, а в качестве предикторов $\xi^{(1)}$ — окружность ее туловища и $\xi^{(2)}$ — длину от хвоста до холки. Ставится задача определения регрессионной зависимости

$$y = f(x^{(1)}, x^{(2)}) = \mathbf{E}(\eta \mid \xi^{(1)} = x^{(1)}, \xi^{(2)} = x^{(2)})$$

---

[1]Исходные данные примера заимствованы у А. Я. Боярского и публикуются с его любезного разрешения.

по результатам контрольных замеров $\{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1,2,\dots,20}$ 20 коров.

Были подвергнуты расчету и сравнительному анализу три варианта параметризации модели:

вариант 1 (линейный): $f(x^{(1)}, x^{(2)}) = \theta_{10} + \theta_{11} x^{(1)} + \theta_{12} x^{(2)}$;

вариант 2 (степенной): $f(x^{(1)}, x^{(2)}) = \theta_{20} (x^{(1)})^{\theta_{21}} (x^{(2)})^{\theta_{22}}$;

вариант 3 (учитывающий содержательный смысл задачи): $f(x^{(1)}, x^{(2)}) = \theta (x^{(1)})^2 x^{(2)}$.

Происхождение варианта 3 легко объяснить. Для этого следует представить себе приближенно тушу коровы в форме цилиндра с длиной образующей, равной $x^{(2)}$, и радиусом основания, равным $x^{(1)}/2\pi$. Используя формулу вычисления объема цилиндра и пропорциональную зависимость между весом и объемом цилиндра, получаем зависимость вида

$$\eta = \theta (x^{(1)})^2 x^{(2)} + \varepsilon,$$

где остаточная компонента ε отражает специфику формы туловища каждой конкретной коровы.

Для проверки работоспособности всех трех вариантов моделей были проведены два цикла расчетов по методу наименьших квадратов (см. гл. 7). Вначале были оценены коэффициенты θ моделей *по всем 20 наблюдениям* и подсчитаны (*по тем же 20 наблюдениям*) характеристики «качества» моделей: множественный коэффициент корреляции $R_{y.(x^{(1)}, x^{(2)})}$ (см. формулу (1.24′)) и остаточные среднеквадратические отклонения

$$s_{\text{ост}} = \sqrt{\widehat{\Delta}'_n (f)} = \sqrt{\frac{1}{n-m} \sum_{i=1}^{n} (y_i - f(x_i^{(1)}, x_i^{(2)}; \Theta))^2} \qquad (6.3)$$

(здесь m — размерность оцениваемого векторного параметра Θ, а $\widehat{\Delta}'_n (f)$ отличается от выборочного критерия адекватности $\widehat{\Delta}_n (f)$ лишь множителем $\frac{n}{n-m}$, см. формулу (5.4′)).

Результаты первого цикла расчетов приведены в гр. 2, 3 и 4 табл. 6.1. Из них как будто следует, что формально-аппроксимационные варианты 1 и 2 оказались несколько точнее варианта 3, выбранного с учетом содержательного смысла задачи.

Однако «благополучие» моделей 1 и 2 лишь кажущееся, что и выявляется в ходе второго цикла вычислений, когда имеющаяся выборка из 20 наблюдений была разбита на две: первая, состоящая из 10 тяжелых коров, была использована для оценки параметров по методу наименьших квадратов (такие выбор-

ки называют *обучающими*), а вторая, состоящая из 10 легких коров, была использована для оценки величины выборочного критерия адекватности $\widehat{\Delta'_n}(f)$ (такие выборки называют *экзаменующими*[1]). Из гр. 5 и 6 табл. 6.1 мы видим, что формально-аппроксимационные варианты моделей не выдержали «экзамен» на устойчивость (сравните значения коэффициентов θ в гр. 2 и 5), и, кроме того, дают явно худшую точность при их использовании в задачах *экстраполяции* (сравните первые две строки с третьей в гр. 6).

Таблица 6.1

Номер варианта модели	Результаты расчетов по всем 20 наблюдениям			Оценки коэффициентов моделей θ по 10 тяжелым коровам	Оценки $\sqrt{\widehat{\Delta'_{10}}(f)}$ по 10 легким коровам
	Оценки коэффициентов θ	$\hat{R}_{y.(x^{(1)},\, x^{(2)})}$	$\sqrt{\widehat{\Delta'_{20}}(f)}$		
1	2	3	4	5	6
1	$\widehat{\theta_{1.0}} = -984,7$ $\widehat{\theta_{1.1}} = 4,73$ $\widehat{\theta_{1.2}} = 4,70$	0,84	25,9	$\widehat{\theta_{1.0}} = 453,2$ $\widehat{\theta_{1.1}} = 0,62$ $\widehat{\theta_{1.2}} = -0,22$	81
2	$\widehat{\theta_{2.0}} = 0,0011$ $\widehat{\theta_{2.1}} = 1,556$ $\widehat{\theta_{2.2}} = 1,018$	0,85	24,5	$\widehat{\theta_{2.0}} = 266,4$ $\widehat{\theta_{2.1}} = 0,203$ $\widehat{\theta_{2.2}} = -0,072$	79
3	$\hat{\theta} = 1,13 \cdot 10^{-4}$	0,83	26,6	$\hat{\theta} = 1,11 \cdot 10^{-4}$	28

Этот пример убедительно демонстрирует, помимо предпочтительности экстраполяционных и «устойчивых» свойств модели 3, что *не следует гнаться за чрезмерной сложностью модели*, ориентируясь при этом на минимизацию выборочного кри-

[1]Подробнее о разбиении выборки на обучающую и экзаменующую см. гл. 11.

терия адекватности $\widehat{\Delta}_n(f)$, когда и оценки неизвестных значений параметров Θ модели и значение критерия $\widehat{\Delta}_n(f)$ вычисляются на основании *одной и той же выборки*. Несостоятельность подобного подхода можно пояснить и теоретически: в соответствии с известным в математическом анализе результатом для любой заданной системы из n точек плоскости $(x_1, y_1), \ldots, (x_n, y_n)$ (с неповторяющимися абсциссами) можно подобрать такой алгебраический полином степени $n - 1$, который пройдет через все точки этой системы. А значит, увеличивая число параметров в параметрическом семействе функций, задающем класс допустимых решений, мы можем добиться «идеальной точности» в смысле *нулевого* значения критерия $\widehat{\Delta}_n(f)$.

На том, чего и как надо добиваться в действительности, мы подробнее остановимся в § 6.2, 6.3 и в гл. 11.

6.2. Предварительный анализ геометрической структуры исходных данных

При выяснении вопроса о параметрическом виде исследуемой зависимости, как правило, идут от простого к сложному. Простейшей же аппроксимацией неизвестной функции регрессии $f(X) = \mathbf{E}(\eta \mid \xi = X)$ является, естественно, *линейная модель*, т. е. функция вида

$$f_a(X) = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)}. \tag{6.4}$$

В предыдущей главе (см. п. 5.1) уже упоминалось, что если анализируемые переменные $(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)}; \eta)$ подчиняются $(p + 1)$-мерному нормальному закону распределения, то истинная функция $f(X)$ регрессии η по $\xi^{(1)}, \ldots, \xi^{(p)}$ принадлежит классу линейных (по $x^{(k)}$, $k = 1, 2, \ldots, p$) функций (6.4). Однако статистическая проверка многомерной нормальности изучаемой векторной случайной величины относится к задачам, до сих пор плохо оснащенным достаточно эффективным инструментарием для их решения (см. сноску к с. 152 [14]). К тому же возможны ситуации, когда анализируемый многомерный признак $(\xi^{(1)}, \ldots, \xi^{(p)}; \eta)$ не является нормальным, но в то же время регрессия η по $(\xi^{(1)}, \ldots, \xi^{(p)})$ линейна.

Поэтому при *предварительном* анализе характера исследуемых зависимостей (т. е. до проведения вычислительных процедур по оценке неизвестных значений параметров, входящих в гипотетичные уравнения связей) ограничиваются некоторыми приближенными эвристическими приемами, связанны-

ми в основном с изучением «геометрии» парных корреляционных полей.

6.2.1. Содержание геометрического анализа парных корреляционных полей. Под *корреляционным полем* переменных (u, v) понимается графическое представление имеющихся измерений (u_1, v_1), (u_2, v_2),..., (u_n, v_n) этих переменных в плоскости (u,v). Мы уже неоднократно имели дело с корреляционными полями (см. рис. В.2, В.4—В.7, 1.1, 5.1, 6.1).

Анализ парных корреляционных полей состоит обычно в следующем:

а) построение на основании имеющихся исходных данных вида (6.1) корреляционных полей для всевозможных пар переменных вида $(x^{(j)}, x^{(k)})$ и $(x^{(j)}, y)$, отобранных из набора всех $p+1$ исследуемых признаков $(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; y)$; всего таких пар будет, очевидно, $p(p+1)/2$, однако процесс этот легко автоматизируется с помощью средств современных ЭВМ;

б) визуальное прослеживание характера вытянутости каждого корреляционного поля: эллипсоидально-линейное (см. рис. В.6), нелинейно-монотонное (см. рис. 6.1), с наличием одного или нескольких экстремумов (см. рис. В.4) и т. п.;

в) изучение поведения условных средних значений результирующего показателя при изменении величины переменной, откладываемой по оси абсцисс и играющей роль предикторной (см. рис. В.2); для этого (если значения предикторной переменной *неконтролируемы* в ходе наблюдения или эксперимента) предварительно разбивают диапазон значений объясняющей переменной *на интервалы группирования* (см. [14], п. 5.4.2) и подсчитывают средние значения ординат тех точек-наблюдений, которые попали в общий интервал группирования.

В результате такого анализа обычно получают формулировку нескольких рабочих гипотез об общем виде искомой зависимости, окончательная проверка которых и выбор наиболее адекватной из них осуществляются (при отсутствии априорных сведений содержательного характера) с помощью соответствующих *математико-статистических* методов. Описание наиболее эффективных, с нашей точки зрения, приемов такого типа приводится в § 6.3. Здесь же остановимся на двух вспомогательных приемах, которые полезно использовать при геометрическом анализе парных корреляционных полей.

6.2.2. Учет и формализация «гладких» свойств искомой функции регрессии. Выше упоминалось, что чрезмерное усложнение класса допустимых решений **F** и, в частности, завышение порядка аппроксимирующего регрессионного полинома (в по-

гоне за снижением значения выборочного критерия адекватности $\widehat{\Delta}_n(f_a)$) может привести к неоправданному усложнению вида искомой функции $f(x) = \mathbf{E}(\eta|\xi = x)$, когда случайные отклонения исходных ($(x_i, y_i), i = 1, \ldots, n$) или условно осредненных по η ($(x_k^0, \overline{y}_k), k = \overline{1,s}$) точек неправильно истолковываются как определенные закономерности в поведении регрессионной кривой. На рис. 6.2 представлен наглядный пример такого переусложнения, когда, располагая таблицей исходных данных вида табл. 6.2

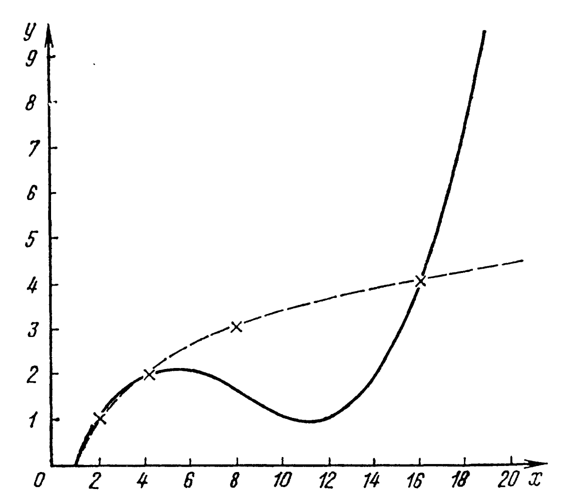


Рис. 6.2. Аппроксимация регрессионной функции $y = \log_2 x$ (пунктирная кривая) с помощью полинома 3-го порядка

Таблица 6.2

k	1	2	3	4
x_k^0	1	2	4	16
$y_{\text{ср}}(x_k^0)$	0	1	2	4

и подбирая аппроксимирующий полином

$$f_a(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3,$$

проходящий через все заданные точки $(x_k^0,\ y_{ср}(x_k^0))$, $k = 1, \ldots, 4$*, приходят к необоснованному нарушению гладкости неизвестной истинной функции регрессии $f(x) = \log_2 x$. Из рис. 6.2 мы видим, что это нарушение гладкости уводит нас достаточно далеко от истины как для значений x, расположенных внутри отрезка [5; 15], так и при $x \geqslant 17$. Поэтому не следует забывать, что если истинный общий вид функции регрессии нам не известен и мы вынуждены ее формально аппроксимировать (например, алгебраическим полиномом), то *всякая интерполяция и тем более экстраполяция*[1] *построенной нами аппроксимационной функции регрессии является, строго говоря, действием, теоретически не обоснованным.* Приведенный пример предупреждает нас о необходимости быть очень осторожными при истолковании и применении регрессионных уравнений, не использующих специальные сведения об изучаемом процессе или явлении.

Интуитивные соображения относительно соблюдения необходимых свойств гладкости, высказываемые при выборе общего вида функции регрессии $f(x)$, могут быть формализованы с помощью так называемых функционалов гладкости $L(f)$. Эти функционалы[2] устроены таким образом, что чем более гладкой, более плавной является функция $f(x)$, тем меньшее числовое значение они принимают. Нетрудно показать, что к такого рода функционалам относятся функционалы вида

$$L_1(f) = \max_{x \in \mathbf{X}} |f'(x)|; \quad L_2(f) = \int_{\mathbf{X}} (f'(x))^{3/2}\,\mathrm{d}x;$$

$$L_3(f) = \int_{\mathbf{X}} f''(x)\,\mathrm{d}x.$$

Приведем пример, в котором выбор функционала гладкости и требование его минимизации поддаются четкой физической интерпретации. Формально задача выглядит так. Рассматри-

*Очевидно, подбор числовых значений θ_j ($j = 0, 1, 2, 3$) осуществляется с помощью решения системы уравнений $y_{ср}(x_k^0) = \theta_0 + \theta_1 x_k^0 + \theta_2 (x_k^0)^2 + \theta_3 (x_k^0)^3$, $k = 1, 2, 3, 4$, которое в данном случае дает: $\widehat{\theta}_0 = -1{,}410$; $\widehat{\theta}_1 = 1{,}633$, $\widehat{\theta}_2 = -0{,}233$; $\widehat{\theta}_3 = 0{,}0095$.

[1] *Интерполяция* — восстановление значений функции (в данном случае — функции регрессии) по значениям аргумента, расположенным *внутри* статистически обследованной области предикторных переменных. *Экстраполяция* — восстановление значений функции регрессии по значениям аргумента, расположенным *вне* статистически обследованного диапазона предикторной переменной.

[2] Функционал $L(f)$ ставит в соответствие каждой заданной *функции* $f(x)$ некоторое *число* $L(f)$.

вается парная регрессионная схема типа B (см. § B.5)

$$\eta = f(x) + \varepsilon(x)$$

с *известной* величиной дисперсии остаточной случайной компоненты $\varepsilon(x)$: $\mathbf{D}\,\varepsilon(x) = \sigma^2$.
Имеются результаты наблюдений $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Требуется определить такую выборочную аппроксимацию $\widehat{f}_a(x)$ функции регрессии $f(x)$, для которой одновременно выполнялись бы условия

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}_a(x_i))^2 = \sigma^2; \tag{6.5}$$

$$\int_{-\infty}^{\infty} \widehat{f}''_a(x)\,\mathrm{d}x = \min_{f_a} \int_{-\infty}^{\infty} f''_a(x)\,\mathrm{d}x.$$

Другими словами, из всех функций, для которых остаточная дисперсия равнялась бы заданной величине σ^2, мы должны выбрать наиболее гладкую в смысле минимизации функционала гладкости $L_3(f)$. Можно привести пример простой физической интерпретации формальной модели (6.5): если мы рассмотрим бесконечную тонкую гибкую рейку, закрепленную в точках $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, но закрепленную не «намертво», а с помощью пружинок заданной силы (пропорциональной σ^{-2}), то эта рейка изогнется как раз по кривой $y = \widehat{f}_a(x)$, определяемой соотношениями (6.5).

Более подробно о результатах, относящихся к решению задач типа (6.5), см. [123].

6.2.3. Некоторые вспомогательные преобразования, линеаризующие исследуемую парную зависимость. Часто при рассмотрении парных корреляционных полей ни линейная, ни полиномиальная регрессия не дают желаемой точности приближения. В этих случаях приходится обращаться к другим видам зависимостей: гиперболической, степенной, показательной и др. Покажем, что в ряде ситуаций эти зависимости оказываются не менее удобными, чем линейная, поскольку легко к ней сводятся.

Так, в примере В.2 при исследовании зависимости между долговечностью образцов N и величиной соответствующего эксплуатационного напряжения v роль зависимой переменной играет величина $\eta = \lg(N - N_0)$, а аргумента — $x = \lg v$. Поэтому, исследуя линейную зависимость между η и x, мы в действительности исследуем соотношение степенного вида между *исходными* переменными N и v, а именно зависимость

вида $N - N_0 = A_1 \cdot v^{b_1}$, где N_0, A_1 и b_1 — некоторые постоянные величины, две из которых (A_1 и b_1) подбираются с помощью метода наименьших квадратов (см. гл. 7).

Перейти от степенной зависимости к линейной нам позволило логарифмическое преобразование переменных.

Какие же функциональные зависимости поддаются линеаризации, каковы их основные свойства, геометрическая интерпретация? С помощью каких преобразований переменных сводятся они к линейному виду?

Итак, пусть η' и x' — исходные переменные (соответственно функция и аргумент), связь между которыми подлежит статистическому исследованию. И пусть между ними существует зависимость

$$\eta' = f(x') + \varepsilon'(\eta', x'), \tag{6.6}$$

где η' — случайная зависимая переменная, x' — аргумент (случайный или неслучайный), $f(x')$ — некоторая функция от x', а $\varepsilon'(\eta', x')$ — так называемая остаточная случайная величина, характеризующая разброс случайных значений η' около функции $f(x')$, которая в самом общем случае может зависеть (стохастически) и от η', и от x'. Поскольку математическое ожидание остаточной случайной величины $\varepsilon'(\eta', x')$ при любых η' и x' равно нулю, то из (6.6) следует, что условное среднее $\mathbf{E}(\eta'|x') = y'_{\text{ср}}(x')$ связано с x' соотношением $y'_{\text{ср}}(x') = f(x')$.

Рассмотрим некоторые наиболее распространенные типы зависимостей $f(x')$ и способы их линеаризации.

Зависимости гиперболического типа (*рис. 6.3, 6.4, 6.5*)

1) $y_{\text{ср}} = a' + \dfrac{b}{x'} = \dfrac{a'x' + b}{x'} \quad (0 < x' < \infty)$.

Этот тип кривых (рис. 6.3) характеризуется двумя асимптотами (прямыми, к которым график функции неограниченно приближается, не достигая их): горизонтальной $y = a'$ и вертикальной $x' = 0$, а также параметром искривления b. С помощью преобразования независимой переменной $x = 1/x'$ (т. е. перехода к новому аргументу) эта зависимость приводится к линейному виду $y = a' + bx$;

2) $y'_{\text{ср}} = \dfrac{1}{a' + bx} \quad \left(-\dfrac{a'}{b} < x < \infty\right)$.

В этом случае имеются две асимптоты: $y' = 0$ и $x = -a'/b$ (рис. 6.4). Параметр, характеризующий искривление, равен $1/b$. Зависимость линеаризуется с помощью перехода к

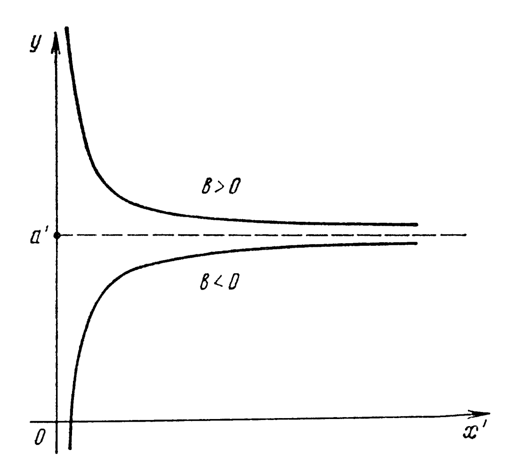


Рис. 6.3. График гиперболической зависимости вида $y = a' + b/x'$

новой зависимой переменной $\eta = 1/\eta'$ (для выборочных значений $y_i = 1/y'_i$);

$$3)\ y'_{\text{ср}} = \frac{x'}{a'\,x' + b} = \frac{1}{a' + b/x'} \quad \left(-\frac{b}{a'} < x < \infty\right).$$

Рассматриваемые кривые (рис. 6.5) имеют горизонтальную асимптоту $y' = 1/a'$, вертикальную асимптоту $x' = -b/a'$ и характеристику искривления, равную $-b/a'^2$. С помощью

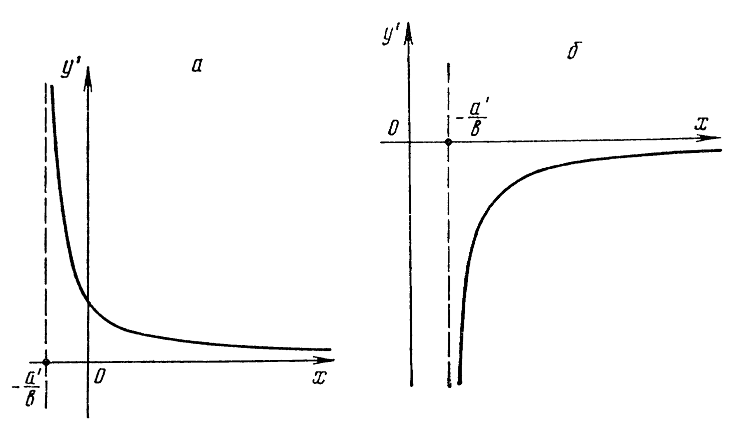


Рис. 6.4. График гиперболической зависимости вида $y' = 1/(a' + bx)$:
а) случай $b > 0$, $a' < 0$; б) случай $b < 0$, $a' < 0$

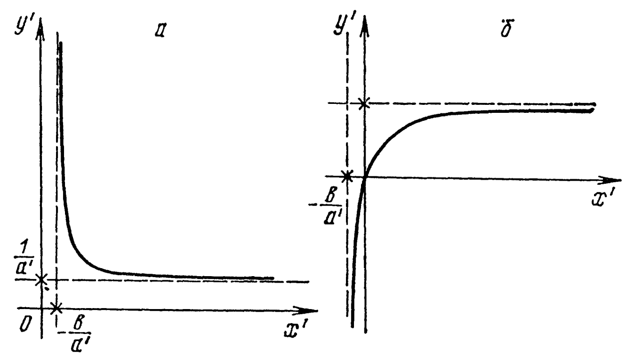


Рис. 6.5. График гиперболической зависимости вида $y' = x'/(a'x'+b)$:
а) случай «положительного» исправления $(-b/a'^2 > 0)$; б) случай «отрицательного» исправления $(-b/a'^2 < 0)$

перехода к переменным $\eta = 1/\eta'$ и $x = 1/x'$ кривые приводятся к линейному виду.

Зависимости показательного типа (рис. 6.6, 6.7, 6.8)

1) $y'_{ср} = A\, e^{bx}$ $(-\infty < x < \infty)$.

Кривые (рис. 6.6) проходят через точку $(0, A)$, причем ось x является их горизонтальной асимптотой. Если вместо η' (соответственно y'_i) в качестве зависимой переменной рассмотреть величину $\eta = \ln \eta'$ (соответственно $y_i = \ln y'_i$), то данная зависимость преобразуется к линейному виду $y_{ср} = a' + bx$, в котором $a' = \ln A$;

2) $y'_{ср} = Ae^{b/x'}$ $(0 < x' < \infty)$.

При $b > 0$ кривая (рис. 6.7, *а*) имеет горизонтальную асимптоту $y' = A$ и вертикальную асимптоту $x' = 0$. При $b < 0$

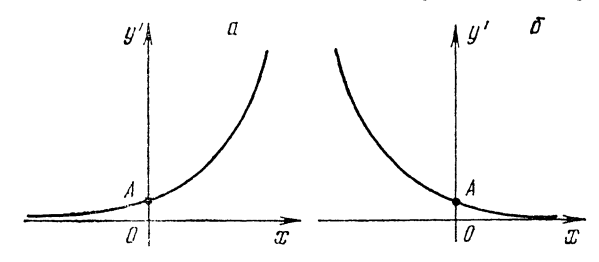


Рис. 6.6. График показательной (экспоненциальной) зависимости вида $y' = Ae^{bx}$: а) случай $b > 0$; б) случай $b < 0$

(рис. 6.7, *б*) кривая проходит через начало координат, имеет так называемую «точку перегиба» ($-b/2$, A/e^2) и горизонтальную асимптоту $y' = A$. Переход к переменным $\eta = \ln \eta'$ (соответственно $y_i = \ln y_i'$) и $x_i = 1/x_i'$ позволяет линеаризовать и эту зависимость, причем в преобразованном виде $y_{ср} = a' + bx$, параметр $a' = \ln A$;

3) $y'_{ср} = \dfrac{1}{a' + be^{-x'}}$ $(-\infty < x' < \infty)$.

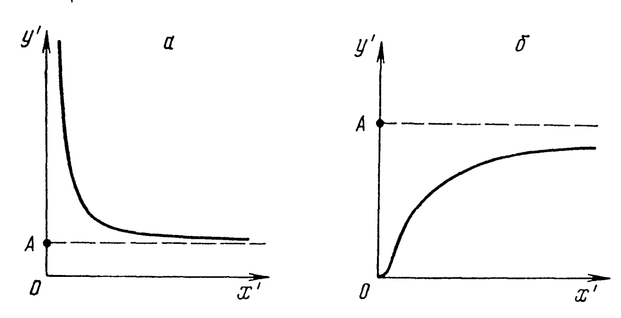


Рис. 6.7. График показательной (экспоненциальной) зависимости вида $y' = Ae^{b/x'}$: а) случай $b > 0$; б) случай $b < 0$

Частный случай так называемой «логистической» кривой показан на рис. 6.8. Кривая имеет две горизонтальные асимптоты $y' = 0$ и $y' = 1/a'$ и «точку перегиба» ($\ln (b/a')$, $1/2a$). Линеаризация этой зависимости производится с помощью перехода к новым переменным $\eta = 1/\eta'$ (соответственно $y_i = 1/y_i'$) и $x = e^{-x'}$.

Зависимости степенного типа (рис. 6.9)

$y'_{ср} = Ax'^{b}$ $(0 \leqslant x' < \infty)$.

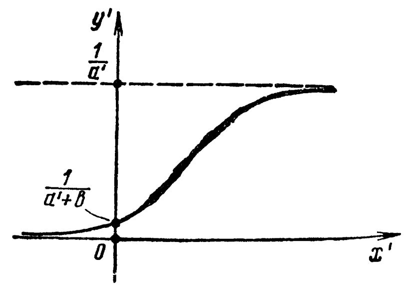


Рис. 6.8. График логистической кривой, описываемой уравнением вида $y' = 1/(a' + be^{-x'})$

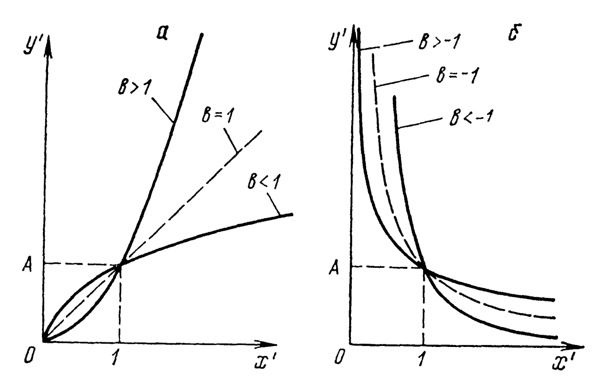


Рис. 6.9. График степенной зависимости вида $y'=A(x')^b$:
а) случай $b>0$; б) случай $b<0$

Все кривые на рисунке проходят через точку (1, A), причем если $b > 0$, то они проходят еще и через начало координат — точку (0, 0), а если $b < 0$, то координатные оси являются одновременно асимптотами. Перейдя к новым переменным $\eta = \ln \eta'$ (соответственно $y_i = \ln y_i'$) и $x = \ln x'$, мы преобразуем исследуемую зависимость к линейному виду.

Зависимости логарифмического типа (рис. 6.10)

$$y_{ср} = a' + b \cdot \ln x' \quad (0 < x < \infty).$$

Кривые на рисунке проходят через точку (1, a') и имеют в качестве вертикальной асимптоты ось y (т. е. $x' = 0$). Переход к линейному виду зависимости осуществляется с помощью логарифмического преобразования аргумента: $x = \ln x'$.

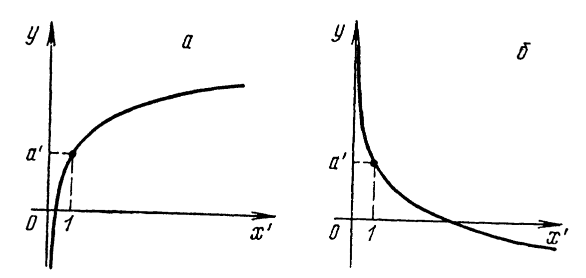


Рис. 6.10. График логарифмической зависимости вида $y=a'+b \cdot \ln x'$:
а) случай $b>0$; б) случай $b<0$

З а м е ч а н и е. Линеаризация связей с помощью преобразования исследуемых переменных имеет недостаток. Оценки параметров a' и b, полученные затем (после линеаризации) с помощью метода наименьших квадратов, на самом деле не минимизируют сумму квадратов отклонений $\widehat{\Delta}_n (a', b) = \frac{1}{n} \sum_{i=1}^{n} (y'_i - f(x'_i))^2$ для *исходных* переменных η' и x'. Они лишь минимизируют сумму квадратов отклонений *преобразованных* значений зависимой переменной y_i от соответствующей регрессионной прямой $y = a' + bx$, т. е. квадратичную форму

$$\widehat{\Delta}_n^{(\text{пр})} (a', b) = \frac{1}{n} \sum_{i=1}^{n} (y_i - a' - bx_i)^2,$$

а это не одно и то же. Предлагается поэтому производить определенную «доводку», уточнение оценок неизвестных значений параметров, полученных с помощью линеаризации связей [10, с. 172].

6.3. Математико-статистические методы в задаче параметризации модели регрессии

6.3.1. Компромисс между сложностью регрессионной модели и точностью ее оценивания[1]. Из общих результатов математической статистики, относящихся к анализу точности оценивания исследуемой модели при ограниченных объемах выборки, следует, что с увеличением сложности модели (например, размерности неизвестного векторного параметра Θ, участвующего в ее уравнении) точность оценивания падает. Мы с этим уже сталкивались, например, при анализе точности оценивания частных и множественных коэффициентов корреляции (см. п.1.2.3, 1.3.3, а также формулы (1.34), (1.34')). Об этом же свидетельствуют и результаты, приведенные в гл. 11. Это означает, в частности, что в ситуациях, когда исследователь располагает лишь ограниченной исходной выборочной информацией, он вынужден искать компромисс между степенью общности привлекаемого класса допустимых решений **F** и точностью оценивания, которой возможно при этом добиться.

Перед тем как изложить общую схему, в рамках которой можно математически ставить и решать задачу достижения та-

---

[1]В изложении материала п. 6.3.1, связанного с понятием *«емкости»* (сложности) класса допустимых решений **F** и с методом *структурной минимизации* критерия адекватности, участвовал В. Н. Вапник.

кого компромисса (метод структурной минимизации критерия адекватности [34]), поясним эту идею на следующем полуэвристическом приеме решения одной частной задачи.

Определение оптимального числа предикторов в модели линейной множественной регрессии. Пусть мы строим линейную множественную регрессию результирующего показателя η по предикторам $(x^{(1)}, x^{(2)}, \ldots, x^{(p)})$, используя для этого выборку *ограниченного* объема

$$(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)};\ y_i)_{i=1, 2, \ldots, n}, \tag{6.7}$$

причем величины p и n — одного порядка (но $p < n - 1$) Очевидно, в данном случае сложность модели будет определяться числом включенных в нее предикторов. Нужно ли для достижения максимальной точности в задаче восстановления неизвестных значений результирующего показателя η по значениям предикторов включать в модель *все* предикторные переменные, а если не все, то *сколько* и *какие именно*?

С одной стороны, мы уже знаем (см. (1.30)), что присоединение каждой новой предсказывающей переменной *может только увеличить* величину множественного коэффициента корреляции R между результирующим показателем η и предикторами и, следовательно, уменьшить ошибку в предсказании $\eta(X)$ (см. (1.26)). С другой стороны, нам известны не *точные* значения теоретических характеристик R, участвующих в (1.26) — (1.30), а лишь их выборочные аналоги — *статистические оценки* $\widehat{R}$.

Поэтому естественно было бы добиваться максимизации не R^2, а *нижней доверительной границы* $(R^2)_P^{\min}$ для истинного значения коэффициента детерминации R^2 (при заданной доверительной вероятности P). Если принять приближенное допущение, что $(R^2)_P^{\min}$ меньше точечной оценки $\widehat{R}^2$ на величину, пропорциональную среднеквадратической ошибке $\sigma_{\widehat{R}} = \sqrt{D\widehat{R}^2}$ (множитель пропорциональности $\lambda(P)$, конечно, зависит от заданной величины доверительной вероятности P), и воспользоваться приближенной формулой (1.34′), то получаем следующую формулу для определения $(R^2)_P^{\min}$:

$$(R^2)_P^{\min} \approx \widehat{R}^2 - \lambda(P)\cdot \frac{2p(n-p-1)}{(n-1)(n^2-1)}(1-\widehat{R}^2). \tag{6.8}$$

Опираясь на (6.8), можно предложить следующую процедуру определения оптимального состава и числа предикторов модели множественной линейной регрессии.

Последовательно для каждого $k=1, 2, \ldots, p$ с использованием формул (1.27), (1.28), (1.35′) рассчитывается величина

$$\widehat{R}^2(k) = \max_{i_1, i_2, \ldots, i_k} \widehat{R}^2_{\eta \cdot x^{(i_1)} x^{(i_2)} \ldots x^{(i_k)}}, \qquad (6.9)$$

а затем по формуле (6.8) — величина $(R^2(k))_P^{\min}$. Тем самым для каждой заданной размерности k модели уже выявлен оптимальный состав предикторов: это тот набор $(x^{(i_1^*)}, x^{(i_2^*)}, \ldots, x^{(i_k^*)})$, на котором достигается максимум правой части (6.9).

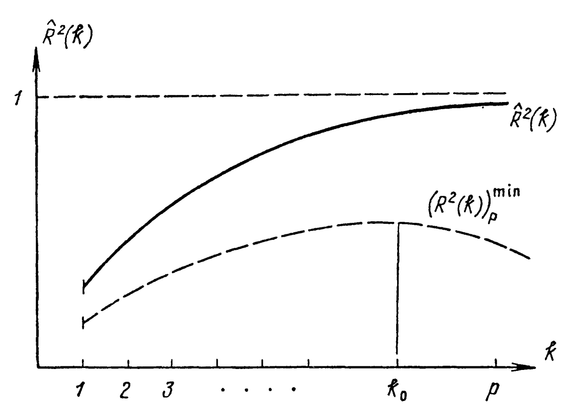


Рис. 6.11. Зависимость нижней доверительной границы коэффициента детерминации от числа предикторов (пунктирная кривая)

На рис. 6.11 представлены схематические графики величин $\widehat{R}^2(k)$ и $(R^2(k))_P^{\min}$ как функций от k.

В качестве оптимального числа предикторов, включаемых в модель, естественно взять то значение k_0, для которого величина $(R^2(k))_P^{\min}$ максимальна.

Метод структурной минимизации критерия адекватности. Опишем теперь, следуя [34], общую схему, в рамках которой решается задача выбора оптимальной сложности параметрического семейства $\{f(X; \Theta)\}$, используемого в качестве класса допустимых решений **F**, в зависимости от объема n и геометрической структуры исходных данных (6.7). Общая логика, на которой построена эта схема, та же, что и логика решения предыдущей задачи: и та, и другая опираются на умение дать гарантированную оценку оптимизируемой *теоретической* характерис-

тики (в общей схеме — оценку сверху для теоретического критерия адекватности $\Delta(\Theta) = \Delta(f(X, \Theta))$ по значению соответствующей *выборочной* характеристики (в общей схеме — по $\widehat{\Delta}(\widehat{\Theta}) = \widehat{\Delta}_n(f(X; \widehat{\Theta}))$. Однако в условиях полного отсутствия какой бы то ни было априорной информации о характере совместного распределения $p(X, y)$ исследуемых переменных (ξ, η) никаких надежных заключений о величине $\Delta(\Theta)$ по значению $\widehat{\Delta}(\widehat{\Theta})$ сделать нельзя. Минимальная информация такого рода, используемая в описываемом методе, состоит в том, чтобы для некоторого $q > 1$ знать величину λ_q, определяющую неравенство

$$\sup_{\Theta \in \Gamma} \frac{\sqrt[q]{\mathrm{E}(y - f(X;\ \Theta))^{2q}}}{\mathrm{E}(y - f(X;\ \Theta))^2} \leqslant \lambda_q.$$

Ниже в целях упрощения формулировок будем требовать выполнения неравенства для случая $q = 2$:

$$\sup_{\Theta \in \Gamma} \frac{\sqrt{\mathrm{E}(y - f(X;\ \Theta))^{4}}}{\mathrm{E}(y - f(X;\ \Theta))^2} \leqslant \lambda. \tag{6.10}$$

Априорная информация, заданная в терминах неравенства (6.10), является более практически доступной, чем обычно используемая связанная с типом распределения регрессионных остатков. Так, если параметрическое семейство случайных величин

$$\varepsilon(X;\ \Theta) = y - f(X;\ \Theta) \tag{6.11}$$

таково, что каждая случайная величина $\varepsilon(X;\ \Theta)$ распределена по *нормальному закону* (со своими параметрами, зависящими от Θ), то $\lambda = \sqrt{3}$; если семейство (6.11) подчинено *закону Лапласа* [14, п. 6.1.8], то $\lambda = \sqrt{5}$; если же $\varepsilon(X;\ \Theta)$ подчинено равномерному закону распределения, то $\lambda = \sqrt{2}$. Неравенство (6.10), по существу, характеризует особенности поведения «хвостов» в случайной выборке наблюдений (6.7).

В ситуации, когда соблюдается условие (6.10), метод структурной минимизации критерия адекватности может быть построен на основе *емкостных характеристик* класса функций $\{f(X; \Theta)\}$, в котором ведется восстановление регрессии. Ниже мы используем одну из возможных таких характеристик — *емкость* (или *сложность*) класса функций $\{f(X;\ \Theta)\}_{\Theta \in \Gamma}$.

Для определения понятия «емкость класса $\{f(X;\ \Theta)\}_{\Theta\in\Gamma}$» введем множество *индикаторных функций*

$$\{J(X;\ y;\ \Theta,\ \beta)\} = \{\text{sign}\,(y - f(X;\ \Theta) + \beta\}^*$$

на элементах $(X_i,\ y_i)$ выборки (6.7).

Индикаторные функции определяются параметрами Θ, β, где $\Theta \in \Gamma$ — параметры, определяющие $f(X;\ \theta)$, и β — некоторое число из интервала $(-\infty, \infty)$. Каждая индикаторная функция $J(X,\ y;\ \Theta^*,\ \beta^*)$ делит выборку (6.7) на две подвыборки: подмножество пар, на которых индикаторная функция принимает значение $+1$, и подмножество пар, на которых индикаторная функция принимает значение -1. Обозначим $N^\Gamma(X_1,\ y_1;\ X_2,\ y_2;\dots;\ X_n,\ y_n)$ количество различных разделений множества (6.7) на два подмножества с помощью индикаторных функций из класса $\{J(X, y;\ \Theta,\ \beta)\}_{\Theta\in\Gamma}$. Очевидно, что $N^\Gamma(X_1,\ y_1;\ \dots;\ X_n,\ y_n) \leqslant 2^n$.

Определение. Назовем функцию

$$m^\Gamma(n) = \max_{X_1,\, y_1;\ \dots;\ X_n,\, y_n} N^\Gamma(X_1,\ y_1;\ \dots;\ X_n,\ y_n) \qquad (6.12)$$

функцией роста класса $\{f(X;\ \Theta)\}_{\Theta\in\Gamma}$ на выборках вида (6.7). Для функции роста справедливо следующее утверждение.

Утверждение. Определенная соотношением (6.12) функция роста либо тождественно равна 2^n, либо, если для некоторого h это не так, т. е. $m^\Gamma(h+1) \neq 2^{h+1}$, то для $n > h$ справедлива оценка

$$m^\Gamma(n) < \frac{n^h}{h!}.$$

Это утверждение позволяет оценивать функцию роста любого класса $\{f(X;\ \Theta)\}_{\Theta\in\Gamma}$ функций. Для получения соответствующей оценки достаточно указать такие $h+1$ пар, которые не могут быть разеделены всеми 2^{h+1} способами с помощью индикаторных функций $J(X,\ y;\ \Theta,\ \beta)$.

В частности, для функций $f(X;\ \Theta)$, линейных по параметрам Θ, т. е.

$$f(X;\ \Theta) = \sum_{k=1}^{m-1} \theta^{(k)}\cdot\psi_k(X) + \theta_0,$$

---

*Напомним, что функция $\text{sign}\,(z)$ определяется соотношениэм

$$\text{sign}\,(z) = \begin{cases} +1, & \text{если } z \geqslant 0; \\ -1, & \text{если } z < 0. \end{cases}$$

где $\{\psi_1(X), \psi_2(X), \ldots, \psi_{m-1}(X)\}$ — некоторая заданная система известных функций, имеет место оценка

$m^\Gamma(n) < n^m/m!$ $(n > m)$.

Итак, функция роста m^Γ (h) оценивается

$$m^\Gamma(h) = \begin{cases} \text{либо } 2^n; \\ \text{либо} < \dfrac{n^h}{h!}, \text{ если } m^\Gamma(h+1) \neq 2^{h+1}, \ n > h. \end{cases}$$

Определение. Будем говорить, что класс функций $f(X; \Theta)$ имеет *бесконечную емкость*, если соответствующая функция роста тождественно равна 2^n, и имеет *конечную емкость* h, если функция роста оценивается сверху величиной $\frac{n^h}{h!}$ $(n > h)$.

Справедливо утверждение: с вероятностью $1 - \alpha$ одновременно для всех функций $\{f(X; \Theta)\}_{\Theta \in \Gamma}$ имеет место неравенство

$$\Delta(\Theta) < \frac{\widehat{\Delta}_n(\Theta)}{\left[1 - \lambda \sqrt{\dfrac{h[\ln(n/h) + 1] - \ln\alpha}{n}\left(1 - \dfrac{1}{4}\ln\dfrac{\left(\ln\dfrac{n}{h} + 1\right)h - \ln\alpha}{n}\right)}\right]_+} \quad (6.13)$$

где

$$[A]_+ = \begin{cases} A, \text{ если } A \geqslant 0; \\ 0, \text{ если } A < 0. \end{cases}$$

Так как неравенство (6.13) с вероятностью $1 - \alpha$ выполняется одновременно для *всех* функций, то оно справедливо и для функции, минимизирующей эмпирический критерий адекватности. Оценка (6.13), по существу, зависит от относительного (по отношению h) объема выборки n.

Таким образом, в условиях (6.10) для класса $\mathbf{F}$ функций ограниченной емкости по величине эмпирического критерия адекватности $\widehat{\Delta}_n(\Theta)$ удается оценить величину теоретического критерия адекватности $\Delta(\theta)$.

Теперь на основе полученной оценки (6.13) сконструируем *метод структурной минимизации критерия адекватности*.

Пусть на исходном классе $\mathbf{F} = \{f(X; \Theta)\}_{\Theta \in \Gamma}$ *задана структура*

$$\mathbf{F}_1 \subset \mathbf{F}_2 \subset \ldots \subset \mathbf{F}_q, \quad (6.14)$$

т. е. задано минимальное подмножество $\mathbf{F}_1$ элементов из $\mathbf{F}$, затем подмножество элементов $\mathbf{F}_2$, содержащее $\mathbf{F}_1$, и т. д. и, наконец, подмножество $\mathbf{F}_q = \mathbf{F}$, содержащее все элементы класса $\{f(X;\Theta)\}_{\Theta\in\Gamma}$. Итак, подмножества $\mathbf{F}_1,\ldots,\mathbf{F}_j,\ldots,\mathbf{F}_q$, таковы, что с ростом номера j емкость их растет: $h_1 < h_2 < \ldots < h_q$.

На каждом из подмножеств $\mathbf{F}_j$ найдем функцию $f(X;\widehat{\Theta}^j)$, минимизирующую выборочный критерий адекватности. Вычислим для функции $f(X;\widehat{\Theta}^j)$ величину выборочного критерия адекватности $\widehat{\Delta}_n(\widehat{\Theta}^j)$. Очевидно, что с ростом номера j величина $\widehat{\Delta}_n(\widehat{\Theta}^j)$ не возрастает.

Но при фиксированных n и α оценка (6.13) величины теоретического критерия адекватности для функции, минимизирующей выборочный критерий адекватности на элементах структуры $\{\mathbf{F}_j\}_{j=\overline{1,q}}$, достигает своего наименьшего значения не обязательно на подмножестве $\mathbf{F}_q = \mathbf{F}$. Иначе говоря, для фиксированного объема выборки наилучшее приближение к функции регрессии достигается на некотором элементе структуры $\mathbf{F}_{j^*}$. Этот метод назван в [34] *методом структурной минимизации риска* (в нашей терминологии — теоретического критерия адекватности). Для ограниченного объема исходных данных n он позволяет установить компромисс между «сложностью» выбираемой модели регрессии (номером элемента структуры (6.14)— чем больше номер, тем сложнее модель) и качеством приближения к выборочным данным (величиной $\widehat{\Delta}_n(\widehat{\Theta}^j)$), при котором достигается наименьшая гарантированная оценка теоретического критерия адекватности. Можно сказать, что дальнейшее усложнение модели приводит к приближению к имеющемуся эмпирическому материалу, а не к искомой зависимости.

Метод структурной минимизации риска может быть использован для восстановления регрессии в различных классах функций. Применим его для построения полиномиальной регрессии.

Пусть структура $\mathbf{F}_1 \subset \ldots \subset \mathbf{F}_q$ такова, что элемент $\mathbf{F}_j$ содержит полиномы степени $j-1$. В этом случае емкость класса $\mathbf{F}_j$ равна j. И проблема заключается в том, чтобы минимизировать выборочный критерий адекватности в классе полиномов такой степени j^*, чтобы достичь минимума оценки (6.13).

На рис. 6.12 показан пример восстановления полинома пятой степени на отрезке $[-2,2]$. Восстановление проводилось по измерениям функции в 20 случайно взятых точках (крестики). Видно, что кривая *2* лучше приближает истинную регрессию, чем кривая *1*.

На рис. 6.13 приведен пример восстановления неполиномиальной истинной регрессии в классе полиномов по 20 измерениям (крестики).

При решении этих примеров минимизировался упрощенный вариант правой части (6.13), а именно функционал

$$\frac{\widehat{\Delta}_n(\Theta)}{\left[1-\sqrt{\dfrac{h\ln\left(\dfrac{n}{h}+1\right)-\ln\alpha}{n}}\right]_+},$$

где было принято $\ln\alpha = -2$.

6.3.2. Поиск модели, наиболее устойчивой к варьированию состава выборочных данных, на основании которых она оценивается. Идея этого подхода к выбору общего вида исследуемой регрессионной зависимости основана на следующем простом соображении: если общий параметрический вид зависимости $y_{\text{ср}} = f(x^{(1)}, x^{(2)}, \ldots x^{(p)}; \Theta)$ «угадан» правильно, то результаты оценивания $\widehat{\Theta}_1, \widehat{\Theta}_2, \ldots$, параметра Θ по различным подвыборкам выборки $\mathbf{B}_n = \{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; y_i\}_{i=\overline{1,n}}$ будут мало отличаться друг от друга (а следовательно, не сильно будут различаться между собой и соответствующие значения $f(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; \widehat{\Theta}_1)$, $f(x^{(1)}, x^{(2)}, \ldots, x^{(p)}; \widehat{\Theta}_2), \ldots$). И, наоборот, при неудачном выборе общего вида искомой зависимости результаты ее восстановления по различным выборкам, как правило, будут сильно отличаться один от другого.

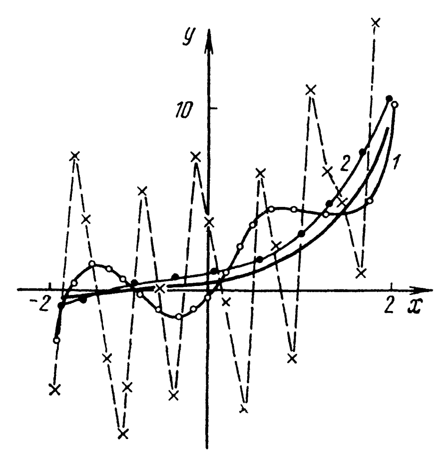


Рис. 6.12. Истинная полиномиальная регрессия (—) и ее аппроксимации: кривая *1* — наилучшее приближение в классе полиномов пятой степени (—○—○—○—); кривая *2* — полученная с помощью алгоритма структурной минимизации критерия адекватности (—•—•—•—)

С проявлением указанного свойства аппроксимационных регрессионных моделей мы уже столкнулись в примере 6.2. Действительно, по данным табл. 6.1 мы видим, что оценки коэффициентов θ_{k0}, θ_{k1} и θ_{k2} ($k = 1,2$) *аппроксимационных вариантов* анализируемой модели (вариантов 1 и

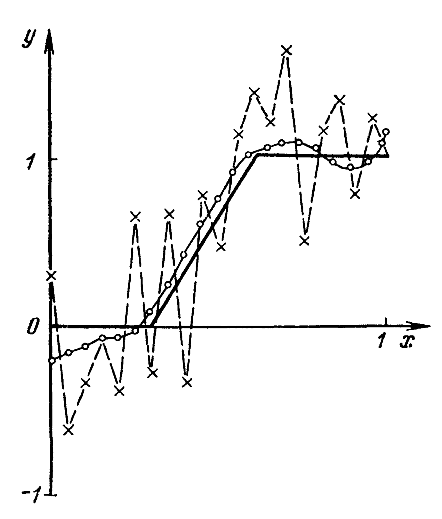


Рис. 6.13. Истинная кусочно-линейная регрессия (—) и ее полиномиальная аппроксимация, полученная с помощью алгоритма структурной минимизации критерия адекватности (—○—○—○—)

2), подсчитанные по различным выборкам (сначала по всей выборке из 20 наблюдений, а затем по ее половине), могут отличаться не только на несколько порядков, но и по знаку (!). В то же время значение оценки коэффициента Θ в модели, общий вид которой *выведен из содержательных соображений* (вариант 3), практически остается одним и тем же при расчете как по всей выборке, так и по ее части.

Предлагаются следующая реализация только что сформулированной идеи и ее экспериментально-вычислительная апробация [1]

Рассмотрим систему **B** подвыборок выборки $\mathbf{B}_n$: $\mathbf{B} = \{b : b \in \mathbf{B}_n\}$.

Пусть на множестве $\overset{*}{\mathbf{X}}$ — области определения исследуемой функции регрессии — задана система линейно-независимых (базисных) функций $\psi_i(X)$, $X \in \overset{*}{\mathbf{X}}$, $i = \overline{1,m}$

Моделью $M_s(X, \widehat{\Theta}(b))$ порядка s для функции $f(X)$, построенной по базису $\{\psi_i(X)\}_{i=1}^m$ и подвыборке $b \in \mathbf{B}$, назовем функцию вида

$$M_s(X, \widehat{\Theta}(b)) = \sum_{i=1}^{s} \widehat{\theta}_i \psi_i(X), \quad s \leqslant m,$$

где коэффициенты $\widehat{\Theta}(b) = (\widehat{\theta}_1, \ldots, \widehat{\theta}_s) = (\widehat{\theta}_1(b), \ldots \widehat{\theta}_s(b))$ являются решением задачи минимизации

$$\min_{\Theta} \sum_{(X_i, y_i) \in b} (y_i - M_s(X_i, \Theta))^2.$$

[1] Описываемая ниже схема изложена на основе результатов и предложений, разработанных В. А. Гусевым (см.: Классификация и аппроксимация экспериментальных данных и надежность прогноза: Автореф. дис. ...канд. физ.-мат. наук. — М., 1982. — В надзаг.: ВЦ АН СССР).

Пусть $\delta \geqslant 0$ — заданное число, а $\mathbf{X}$ — некоторое подмножество из $\overset{*}{\mathbf{X}}$. Назовем множества $b_1 \in \mathbf{B}$ и $b_2 \in \mathbf{B}$ δ_s - эквивалентными ($b_1\ \delta_s\ b_2$), если они удовлетворяют условию

$$|M_s(X, \widehat{\Theta}(b_1)) - M_s(X, \widehat{\Theta}(b_2))| \leqslant \delta,\ X \in \mathbf{X}.$$

Таким образом, δ_s-эквивалентность множеств b_1 и b_2, т. е. подмножеств множества $\mathbf{B}_n$, означает следующее: значение модели $M_s(X, \widehat{\Theta}(b_1))$ функции $f(X)$, определенной по подвыборке b_1, отличается от значения модели $M_s(X, \widehat{\Theta}(b_2))$, определенной по подвыборке b_2, в любой точке X множества $\mathbf{X}$ по модулю на величину, не большую, чем δ. Можно рассматривать δ_s -эквивалентность всей выборки $\mathbf{B}_n$ и ее подвыборок, т. е. сравнивать модель $\overset{\circ}{M}_s(X, \widehat{\Theta}) = M_s(X, \widehat{\Theta}(\mathbf{B}_n))$ с моделями $M_s(X, \widehat{\Theta}(b))$, построенными по отдельным частям выборки $\mathbf{B}_n$.

Рассмотрим такие подвыборки b из $\mathbf{B}_n$, которые содержат ровно α точек, и обозначим их совокупность через $\mathbf{B}^\alpha$, а их число через $m_\alpha = C_n^\alpha$. Далее, определим число $m_\alpha(\delta)$ подвыборок $b \in \mathbf{B}^\alpha$, для которых выполнено условие

$$|\overset{\circ}{M}_s(X, \widehat{\Theta}) - M_s(X, \widehat{\Theta}(b))| \leqslant \delta \text{ при всех } X \in \mathbf{X}.$$

Устойчивость модели $M_s(X, \widehat{\Theta})$ порядка s на множестве $\mathbf{X}$ для заданного δ будем измерять величиной

$$\nu_s(\delta) = \frac{1}{n-s+1} \sum_{\alpha=s}^{n} \frac{m_\alpha(\delta)}{m_\alpha}.$$

Пусть задана последовательность $0 < \delta_1 < \ldots < \delta_t$. Величину $\bar{\nu}_s = \frac{1}{t}\sum_{k=1}^{t} \nu_s(\delta_k)$ назовем *средней устойчивостью модели* на множестве $\mathbf{X}$ для последовательности $(\delta_1, \delta_2, \ldots, \delta_t)$.

Рассмотрим величину

$$\delta_{\max}^s(b) = \max_{X \in \mathbf{X}} |\overset{\circ}{M}_s(X, \widehat{\Theta}) - M_s(X, \widehat{\Theta}(b))|,$$

максимальную по модулю разности моделей на множестве $\mathbf{X}$. Таким образом, можно рассматривать распределение значений величины $\delta_{\max}^s$ на системе подвыборок $\mathbf{B}$. В частности, можно оценить математическое ожидание $\mathbf{E}\delta_{\max}^s$ величины $\delta_{\max}^s$ и квантиль u_P^s порядка P распределения $\delta_{\max}^s$.

Для оценки качества модели можно использовать следующие характеристики:

$\nu_s(\delta)$ — характеристику устойчивости для заданного δ;

$\bar{\nu}_s$ — характеристику средней устойчивости для последовательности $(\delta_1, \delta_2, \ldots, \delta_t)$;

$\mathbf{E}\delta^s_{\max}$ — математическое ожидание величины $\delta^s_{\max}$;

u^s_P — квантиль порядка P распределения величины $\delta^s_{\max}$.

Для наилучшей модели характеристики $\nu_s(\delta)$ и ν_s достигают максимального, а характеристики $\mathbf{E}\delta^s_{\max}$ и u^s_P — минимального значений. Практическая реализация данного подхода, опирающегося на анализ величин $\nu_s(\delta)$, ν_s, $\mathbf{E}\,\delta^s_{\max}$ и u^s_P, требует привлечения ЭВМ и расчета необходимых статистических характеристик этих величин с помощью метода Монте-Карло [14, § 6.3].

Возможна и иная форма реализации данного подхода, не предусматривающая необходимости использования статистического моделирования на ЭВМ. Она основана на анализе критических статистик вида

$$\hat{\gamma}(n_1, n_2, s) = \frac{\dfrac{1}{n_1 - s} \sum\limits_{X_i\, y_i \in b_1} (y_i - M_s(X_i; \hat{\Theta}(b_1))^2}{\dfrac{1}{n_2 - s} \sum\limits_{X_i\, y_i \in b_2} (y_i - M_s(X_i; \hat{\Theta}(b_2))^2}, \qquad (6.15)$$

где b_1 и b_2 — непересекающиеся подвыборки объемов n_1 и n_2 $(n_1 + n_2 < n)$, случайно и независимо извлеченные (без возвращения) из исходной выборки $\mathbf{B}_n$. В частности, в условиях справедливости гипотезы H_0: $\mathbf{E}(\eta | \xi = X) = M_s(X; \Theta)$ случайная величина (6.15) должна подчиняться приблизительно F-распределению с числом степеней свободы числителя и знаменателя $n_1 - s$ и $n_2 - s$ соответственно.

Для статистической проверки этого факта можно воспользоваться сравнением подсчитанного значения статистики $\hat{\gamma}$ с процентной точкой F-распределения (см. табл. П.5). А при достаточно больших объемах n исходных выборок $\mathbf{B}_n$ можно непосредственно проверять факт $\mathbf{F}$-распределенности случайных величин $\hat{\gamma}$ с помощью соответствующих критериев согласия [14, § 11.1]. Для этого, правда, следует образовать целую последовательность подвыборок $b_1, b_2, \ldots, b_N$ из $\mathbf{B}_n$, подсчитать для различных пар b_i, b_j величины (6.15) и применить к ним критерий согласия.

6.3.3. Статистические критерии проверки гипотез об общем виде функции регрессии. Подчеркнем сразу, что описанные ниже критерии проверки справедливости сделанного выбора общего

вида искомой функции регрессии не могут ответить на вопрос: является ли проверяемый гипотетичный вид зависимости наилучшим, единственно верным? Они лишь либо подтверждают факт непротиворечивости проверяемого вида функции регрессии имеющимся у исследователя исходным данным (6.1), либо отвергают обсуждаемую гипотетичную форму зависимости как не соответствующую этим данным.

1. *Общий приближенный критерий, основанный на группированных данных* (или при наличии нескольких наблюдений при каждом фиксированном значении аргумента). Пусть высказана гипотеза об общем виде функции регрессии H_0 : $\mathbf{E}(\eta|\xi = X) = f_a(X; \theta_1, \theta_2, \dots, \theta_m)$ ($f_a(X; \Theta)$ — известная функция, $(\theta_1, \theta_2, \dots, \theta_m) = \Theta$ — неизвестные числовые параметры) и пусть вычислены (например, с помощью метода наименьших квадратов, см. гл. 7) оценки $\widehat{\theta}_1, \widehat{\theta}_2, \dots, \widehat{\theta}_m$ неизвестных параметров, входящих в описание уравнения регрессии. При группировке данных (или при проведении эксперимента) мы должны соблюдать требование, в соответствии с которым число интервалов группирования (или число различных значений аргумента, в которых производились наблюдения) k должно обязательно превосходить число неизвестных параметров m, т. е. $k - m \geqslant 1$.

Если высказанная гипотеза об общем виде зависимости является правильной, то статистика

$$v^2 = \frac{\frac{1}{k-m}\sum_{i=1}^{k} m_i\,|\bar{y}_i - f_a(X_i^0;\ \widehat{\Theta})|^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{m_i} |y_{ij} - \bar{y}_i|^2} \tag{6.16}$$

должна приближенно подчиняться $F(\nu_1, \nu_2)$-распределению с числом степеней свободы числителя $\nu_1 = k - m$ и знаменателя — $\nu_2 = n - k$. Все величины в формуле (6.16) соответствуют ранее введенным обозначениям. В частности, X_i^0 — середина i-го гиперпараллелепипеда группирования (или i-е значение аргумента, в котором было проведено m_i наблюдений); $f_a(X_i^0; \widehat{\Theta})$ — значение гипотетической функции регрессии, вычисленное в точке $X = X_i^0$; $\bar{y}_i$ — условное среднее из ординат, попавших в i-й гиперпараллелепипед группирования (или из ординат, измеренных при i-м фиксированном значении аргумента X_i^0); y_{ij} — j-е по счету значение ординаты из числа попавших в i-й интервал группирования (или из числа измеренных

при i-м фиксированном значении аргумента X_i^0). Легко понять, что числитель в правой части (6.16) характеризует меру рассеивания экспериментальных данных вокруг аппроксимирующей выборочной регрессионной поверхности, а знаменатель — меру рассеивания экспериментальных данных около своих условных выборочных средних $\bar{y}_i$ (т. е. меру, независимую от выбранного вида линии регрессии). Причем и числитель, и знаменатель являются практически независимыми (в некоторых частных случаях — *точно* независимыми) статистическими оценками одной и той же теоретической дисперсии $\sigma^2 = \mathbf{D}(\eta|\xi=X)$.

Соответственно получаем следующее правило проверки гипотезы об общем виде функции регрессии. Задаемся, как обычно, достаточно малым уровнем значимости критерия α (например, $\alpha = 0{,}05$). С помощью табл. П. 5 находим $100\left(1 - \frac{\alpha}{2}\right)$%-ную точку $v^2_{1-\frac{\alpha}{2}}$ и $100\frac{\alpha}{2}$%-ную точку $v^2_{\frac{\alpha}{2}}$ $F(k - m, n - k)$-распределения. Если окажется, что величина v^2, подсчитанная по формуле (6.16), удовлетворяет неравенствам

$$v^2_{1-\frac{\alpha}{2}} < v^2 < v^2_{\frac{\alpha}{2}},$$

то высказанная нами гипотеза об общем виде функции регрессии признается не противоречащей экспериментальным данным (6.1). Если же эти неравенства оказались нарушенными, то гипотеза об общем виде функции регрессии отвергается с уровнем значимости α. При этом если v^2 «слишком мало» (т.е. $v^2 < v^2_{1-\frac{\alpha}{2}}$) то, очевидно, при выборе общего вида регрессии мы неправомерно реагировали на случайные отклонения точек $(X_i^0, \bar{y}_i)$ от истинной функции регрессии и тем самым необоснованно завысили число параметров m, от которых зависит уравнение регрессии. Напротив, если v^2 «слишком велико» (т. е. $v^2 > v^2_{\frac{\alpha}{2}}$), то «гибкость» аппроксимирующей функции регрессии $f_a(X; \Theta)$ следует признать недостаточной, поэтому целесообразно увеличить число неизвестных параметров регрессии (например, повысить порядок аппроксимирующего полинома).

Для случая, когда условная дисперсия зависимой переменной пропорциональна некоторой известной функции аргумента, т. е. $\mathbf{D}\eta(X) = \sigma^2 h^2(X)$, формула (6.16) преобразуется:

$$v'^2 = \frac{\frac{1}{k-m} \cdot \sum_{i=1}^{k} \omega_i m_i |\bar{y}_i - f_a(X_i^0, \hat{\Theta})|}{\frac{1}{n-k} \sum_{i=1}^{k} \omega_i \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}, \qquad (6.16')$$

где

$\omega_i = 1/h^2(X_i^0)$.

Так, в примере В.2: $n = 40$; $k = 4$; $\alpha = 0{,}05$; $m = 2$; дисперсионное отношение v'^2, подсчитанное по формуле (6.16′), равно 1,04, в то время как 5%-ная точка F (2,36)-распределения $v_{0,05}^2 = 3{,}26$. Это свидетельствует о том, что гипотеза о линейном виде регрессионной зависимости в данном случае не противоречит имеющимся в нашем распоряжении экспериментальным данным.

При проверке *линейности* регрессии (так же, впрочем, как и при проверке гипотезы о полиномиальном характере регрессии заданного порядка m) в нормальных схемах зависимостей типа B и C_1 описанный общий критерий является *точным.* При этом в линейном случае статистика v^2, определенная соотношением (6.16), может быть выражена в более удобной форме, не требующей предварительного вычисления выборочной аппроксимирующей функции регрессии, а именно:

$$v^2 = \frac{(n-k)\left(\hat{\rho}_{\eta \cdot \xi}^2 - \hat{r}^2\right)}{(k-2)\left(1 - \hat{\rho}_{\eta \cdot \xi}^2\right)}. \qquad (6.17)$$

Здесь, как и прежде, $\hat{\rho}_{\eta \cdot \xi}$ и $\hat{r}$ — соответственно выборочные корреляционные отношения (η по ξ) и коэффициент корреляции, вычисляемые по формулам (1.16) и (1.8′). Логическая схема использования статистики (6.17) аналогична ранее изложенным критериям: задаются достаточно малым ($0{,}05 \sim 0{,}15$) уровнем значимости α; находят по табл. П.5 100 α%-ную точку v_α^2 распределения $F(k-2,\ n-k)$; сравнивают величину v^2, определенную с помощью (6.17), с процентной точкой v_α^2; если оказывается, что $v^2 > v_\alpha^2$, то гипотезу о линейном виде регрессии считают статистически необоснованной.

Воспользуемся данным критерием для статистической проверки линейности регрессии в примере В.3. Вычисления дают: $\hat{r}^2 = 0{,}429$, $\hat{\rho}_{\eta \cdot \xi}^2 = 0{,}459$, так что $v^2 = 0{,}513$. Принимая во внимание, что величина 5%-ной точки F (4,37)-распределения равна $v_{0,05}^2 = 2{,}63$, делаем вывод о непротиворечивости гипотезы

линейности регрессии и данных нашего эксперимента в данном примере ($0{,}513 < 2{,}63$).

2. *Общий приближенный критерий, основанный на негруппированных данных (при известной величине дисперсии остаточной случайной компоненты).*

Встречаются ситуации, когда в результате предварительных исследований или из других каких-либо соображений нам удается заранее определить величину дисперсии σ^2 остаточной случайной компоненты ε в разложениях вида (В.14) и (В.16) (например, когда ε — ошибка измерения, и нам известны характеристики точности используемого измерительного прибора). В этом случае можно отказаться от стеснительного требования группированности данных и для проверки гипотезы об общем виде функции регрессии воспользоваться фактом $\chi^2(n-m)$-распределенности статистики

$$\gamma^2 = \frac{1}{\sigma^2} \cdot \frac{1}{n-m} \sum_{i=1}^{n} \left(y_i - f_a(X_i; \widehat{\Theta})\right)^2 \tag{6.18}$$

(который имеет место при условии справедливости нашей гипотезы).

Задавшись уровнем значимости критерия α и найдя с помощью табл. П. 4 величины $100\left(1-\frac{\alpha}{2}\right)\%$- и $100\frac{\alpha}{2}\%$-ных точек χ^2-распределения с $n-m$ степенями свободы, соответственно $\chi^2_{1-\frac{\alpha}{2}}(n-m)$ и $\chi^2_{\frac{\alpha}{2}}(n-m)$, проверяем выполнение неравенства

$$\chi^2_{1-\frac{\alpha}{2}}(n-m) < \gamma^2 < \chi^2_{\frac{\alpha}{2}}(n-m),$$

где γ^2 подсчитано по формуле (6.18). Если эти неравенства оказались нарушенными, то от гипотезы H_0 об общем виде функции регрессии следует отказаться. При этом если γ^2 «слишком мало» (т. е. $\gamma^2 \leqslant \chi^2_{1-\frac{\alpha}{2}}(n-m)$), то, очевидно, при выборе общего вида мы неправильно реагировали на случайные отклонения экспериментальных точек (X_i, y_i) и тем самым необоснованно завысили число параметров m, от которых зависит уравнение регрессии.

Напротив, если γ^2 «слишком велико» (т. е. $\gamma^2 \geqslant \chi^2_{\frac{\alpha}{2}}(n-m)$), то «гибкость» аппроксимирующей кривой регрессии $f_a(X; \Theta)$ следует признать недостаточной, поэтому целесообразно увеличить число неизвестных параметров регрессии (например, повысить порядок аппроксимирующего полинома).

Для случая, когда условная дисперсия зависимой переменной (или, что то же, дисперсия остаточной случайной компоненты) не остается постоянной при изменении X, а пропорциональна некоторой известной функции аргумента, т. е. $\mathbf{D}\eta(X) = \sigma^2 h^2(X)$, формула подсчета статистики γ^2 несколько изменится:

$$\gamma^2 = \frac{1}{\sigma^2} \cdot \frac{1}{n-m} \sum_{i=1}^{n} \omega_i (y_i - f_a(X_i; \widehat{\Theta}))^2,$$

где $\omega_i = 1/h^2(X_i)$. В остальном схема проверки гипотезы об общем виде функции регрессии остается той же самой, что и в случае $\mathbf{D}\eta(X) = \sigma^2 = \text{const}$.

3. *Оценка размерности модели регрессии.* Предположим, что неизвестная истинная функция регрессии $f(X)$ представима в виде разложения по заданной системе базисных функций

$$\mathbf{E}(\eta \mid \xi = X) = f(X) = \sum_{j=1}^{m_0} \theta_j, \psi_j(X), \qquad (6.19)$$

а регрессионные остатки ε в моделях (В.14), (В.16) — независимые нормальные случайные величины с нулевым математическим ожиданием и дисперсией σ^2. Параметры m_0, $\Theta = (\theta_1, \ldots, \theta_{m_0})$ и σ^2 не известны исследователю. Величину m_0 будем называть *размерностью модели регрессии.* Рассмотрим два способа оценивания m_0, и, следуя [97], опишем статистические свойства такого оценивания.

Оба способа основаны на величине «подправленного» выборочного критерия адекватности

$$\widehat{\Delta}'_n(\widehat{f}^{(m)}) = \frac{1}{n-m} \sum_{i=1}^{n} (y_i - f^{(m)}(X_i; \widehat{\Theta}))^2, \qquad (6.20)$$

где $\widehat{f}^{(m)} = f^{(m)}(X; \widehat{\Theta}) = \sum_{j=1}^{m} \widehat{\Theta}_j \psi_j(X)$, а $\widehat{\Theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$ — оценки наименьших квадратов параметров Θ (см. гл. 7).

В первом способе в качестве оценки необходимого числа базисных функций рекомендуется брать величину

$$\widehat{m}_0^{(1)} = \min \{m : \widehat{\Delta}'_n(\widehat{f}^{(m-1)}) > \widehat{\Delta}'_n(\widehat{f}^{(m)}), \qquad (6.21)$$

$$\widehat{\Delta}'_n(\widehat{f}^{(m)}) \leqslant \widehat{\Delta}'_n(f^{(m+1)})\}.$$

Во втором способе с помощью критической статистики

$$v^2(1, n-m-2) = \frac{(n-m-2)\left[(n-m) \cdot \widehat{\Delta}'_n(\widehat{f}^{(m)}) - (n-m-1) \cdot \widehat{\Delta}'_n(\widehat{f}^{(m+1)})\right]}{(n-m-1) \cdot \widehat{\Delta}'_n(\widehat{f}^{(m+1)})}$$

которая при $m_0 = m$ и сделанных выше предположениях подчиняется F-распределению с числом степеней свободы числителя, равным 1, и знаменателя, равным $n - m - 2$ [130, с. 133], последовательно для $m = 1,2, \ldots$ проверяется гипотеза $m_0 = m$ и останавливаются на таком наименьшем $\widehat{m}_0^{(2)}$, при котором гипотеза впервые не отвергается.

В [97] выведены асимптотические (по $n \to \infty$) распределения для оценок $\widehat{m}_0^{(1)}$ и $\widehat{m}_0^{(2)}$. Показано, что для $l = 1, 2$:

$$\lim_{n\to\infty} P\{\widehat{m}_0^{(l)} < m_0\} = 0;$$

$$\lim_{n\to\infty} P\{\widehat{m}_0^{(l)} = m_0 + N\} = \lambda (1-\lambda)^N, \quad N = 0, 1, \ldots,$$

где

$$\lambda = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-1}^{1} e^{-\frac{x^2}{2}} dx \approx 0{,}683 & \text{для } \widehat{m}_0^{(1)}; \\ \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-v_\alpha}^{v_\alpha} e^{-\frac{x^2}{2}} dx & \text{для } \widehat{m}_0^{(2)}. \end{cases}$$

В последнем соотношении $v_\alpha^2 = v_\alpha^2 (1, \; n - \widehat{m}_0^{(2)} - 2)$ — $100\,\alpha\%$-ная точка $F(1, n - \widehat{m}_0^{(2)} - 2)$-распределения.

Эти результаты позволяют, в частности, строить асимптотические доверительные интервалы для неизвестной размерности модели регрессии.

Существуют и другие различные способы оценки размерности модели регрессии, применимые при рассмотрении некоторых частных схем[1].

4. *Анализ регрессионных остатков.* Ряд статистических критериев проверки адекватности используемой аппроксимирующей модели регрессии основан на анализе регрессионных остатков (невязок) $\widehat{\varepsilon}(X_i) = y_i - \widehat{f}_a(X_i)$, $i = 1, 2, \ldots, n$. В основе их конструирования — положение, в соответствии с которым правильный выбор модели $f_a(X)$ предопределяет асимптотическую (по $n \to \infty$) независимость остатков $\widehat{\varepsilon}(X_i)$. Поэтому статистическая проверка правильности выбора общего вида

[1]См., например: Боганик Г. Н. Об установлении порядка уравнения параболической регрессии. — Теория вероятностей и ее применения, т. XII, 1967, № 4, с. 718—727.

функции регрессии сводится к проверке статистической независимости остатков, для чего могут быть использованы, например, критерии, описанные в [14, § 11.3]. На этом же основан и критерий определения порядка полиномиальной регрессии и критерии проверки независимости величин $\widehat{f}(X_i)$ и $\widehat{\varepsilon}(X_i)$ [93].

ВЫВОДЫ

1. *Этап параметризации регрессионной модели*, т. е. выбора параметрического семейства функций (класса допустимых решений), в рамках которого производится дальнейший поиск неизвестной функции регрессии, *является одновременно наиболее важным и наименее теоретически обоснованным* этапом регрессионного анализа.

2. Прежде всего исследователь должен сосредоточить свои усилия на *анализе содержательной сущности* искомой статистической зависимости, чтобы максимально использовать имеющиеся априорные сведения о «физическом» механизме изучаемой связи при выборе общего вида функции регрессии.

3. Важную роль в правильном выборе параметрического класса допустимых решений играет *предварительный анализ геометрической структуры совокупности исходных данных* и в первую очередь анализ геометрии парных корреляционных полей, включающий в себя, в частности, учет и формализацию «гладких» свойств искомой функции регрессии, использование вспомогательных линеаризующих преобразований.

4. Сформулированные с помощью содержательного и геометрического анализа рабочие гипотезы об общем виде искомой функции регрессии могут быть проверены с привлечением соответствующих *математико-статистических критериев*. Среди фундаментальных идей, на которых базируются эти статистические критерии, следует выделить: а) идею компромисса между сложностью регрессионной модели («емкостью» класса допустимых решений) и точностью ее оценивания; б) идею поиска модели, наиболее устойчивой к варьированию состава выборочных данных, на основании которых она оценивается; в) идею проверки гипотез об общем виде функции регрессии на базе сравнения выборочных критериев адекватности и исследования статистических свойств получаемых при этом оценок размерности модели.

Глава 7. ОЦЕНИВАНИЕ НЕИЗВЕСТНЫХ ЗНАЧЕНИЙ ПАРАМЕТРОВ, ЛИНЕЙНО ВХОДЯЩИХ В УРАВНЕНИЕ РЕГРЕССИОННОЙ ЗАВИСИМОСТИ

Рассмотрим общую модель линейной (относительно оцениваемых параметров Θ) регрессии в виде

$$y_i = \sum_{k=1}^{p} \psi_k(Z_i)\cdot\theta_k + \varepsilon_i,\ i = 1, 2, \ldots, n, \tag{7.1}$$

где θ_k — неизвестные параметры, которые надо оценить по выборочным данным (Z_i, y_i), $i=1, \ldots, n$; $\{\psi_k(Z)\}_{k=\overline{1,p}}$ — система известных (базисных) функций векторного аргумента Z, по которым разложена неизвестная функция регрессии $f(Z) = \mathbf{E}(\eta|\xi = Z)$, т. е. $f(Z) = \sum_{k=1}^{p}\psi_k(Z)\cdot\theta_k$; ε_i — случайная погрешность. Сделав замену переменных $x_i^{(k)} = \psi_k(Z_i)$ и учитывая ранее принятые обозначения

$$X_i = (x_i^{(1)}, \ldots, x_i^{(p)})';\ \mathbf{X} = (X_1, \ldots, X_n)';$$

$$Y = (y_1, \ldots, y_n)';\ \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)',\ \Theta = (\theta_1, \ldots, \theta_p)'.$$

модель (7.1) можно представить в виде

$$Y = \mathbf{X}\Theta + \boldsymbol{\varepsilon}. \tag{7.2}$$

Вектор X_i будем называть наблюденным значением предикторной переменной (регрессора).

В данной главе рассматриваются различные способы оценки параметра Θ в зависимости от предположений о природе $\mathbf{X}$ и характере распределения $\boldsymbol{\varepsilon}$.

7.1. Метод наименьших квадратов

7.1.1. Мнк-уравнения. Предположим, что распределение вектора $\boldsymbol{\varepsilon}$ не зависит от $\mathbf{X}$ и нормально с нулевым вектором средних и ковариационной матрицей $\Sigma = \sigma^2\mathbf{I}_n$, где σ^2 — неизвестная дисперсия компонент $\boldsymbol{\varepsilon}$, а $\mathbf{I}_n$ — единичная матрица порядка n. Сформулированное условие записывается

$$\boldsymbol{\varepsilon} \in N(\theta_n, \sigma^2\mathbf{I}_n). \tag{7.3}$$

Оценка параметров в модели (7.2), (7.3) проводится с помощью метода наименьших квадратов (мнк), который описан в

[14, п. 8.6.3]. При этом Θ находится из условия минимизации суммы квадратов отклонений наблюденных значений y от их сглаженных (регрессионных) значений, т. е. величины

$$\|Y-\mathbf{X}\Theta\|=\sum_{i=1}^{n}\left(y_i-\sum_{k=1}^{p}x_i^{(k)}\theta_k\right)^2=(Y-\mathbf{X}\Theta)'(Y-\mathbf{X}\Theta). \quad (7.4)$$

Уравнения метода наименьших квадратов, мнк-уравнения, в случае, когда r — ранг $\mathbf{X}$ равен p, имеют решение

$$\widehat{\Theta}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y. \quad (7.5)$$

Если $r<p$, то в ряде случаев легко ввести дополнительные ограничения на параметры $\mathbf{H}\Theta=0$, где ранг $\mathbf{H}$ равен $p-r$. Пусть $\mathbf{G}=\begin{pmatrix}\mathbf{X}\\ \mathbf{H}\end{pmatrix}$, тогда $\mathbf{G}'\mathbf{G}=\mathbf{X}'\mathbf{X}+\mathbf{H}'_{\mathbf{H}}$ имеет размер $(p\times p)$ и ранг p и

$$\widehat{\Theta}=(\mathbf{G}'\mathbf{G})^{-1}\mathbf{X}'Y. \quad (7.6)$$

Другой путь — использование обобщенной обратной матрицы $(\mathbf{X}'\mathbf{X})^-$ [17] для $\mathbf{X}'\mathbf{X}$. В этом случае

$$\widehat{\Theta}=(\mathbf{X}'\mathbf{X})^-\mathbf{X}'Y. \quad (7.7)$$

Подправленная на несмещенность оценка максимального правдоподобия [14, п. 8.6.3] для дисперсии σ^2 задается формулой

$$s^2=\|Y-\mathbf{X}\widehat{\Theta}\|/(n-r), \quad (7.8)$$

где $\|Y-\mathbf{X}\widehat{\Theta}\|$ часто называют *остаточной суммой квадратов* (ОСК).

7.1.2. Свойства мнк-оценок. В случае когда (7.3) имеет место, $\widehat{\Theta}$ является наилучшей несмещенной оценкой Θ, т. е. $\mathbf{E}\widehat{\Theta}=\Theta$, и для всякой другой оценки $\widetilde{\Theta}$ со свойством $\mathbf{E}\widetilde{\Theta}=\Theta$ для произвольного (неслучайного) $(p\times 1)$-вектора C

$$\mathbf{D}(C'\widehat{\Theta})\leqslant\mathbf{D}(C'\widetilde{\Theta}). \quad (7.9)$$

Если дополнительно потребовать, чтобы ранг $\mathbf{X}=p$, то из общей теории мнк-оценок следует, что

$$\widehat{\Theta}\in N(\Theta,\sigma^2(\mathbf{X}'\mathbf{X})^{-1}); \quad (7.10)$$

$$(\widehat{\Theta}-\Theta)'\mathbf{X}'\mathbf{X}(\widehat{\Theta}-\Theta)/\sigma^2=\chi_p^2; \quad (7.11)$$

$\widehat{\Theta}$ не зависит от s^2; (7.12)

$$\|Y-\mathbf{X}\widehat{\Theta}\|/\sigma^2=(n-p)\,s^2/\sigma^2=\chi^2_{n-p}. \qquad (7.13)$$

Если сохранить требование $\mathrm{E}\varepsilon_i = 0$, $\mathrm{E}\varepsilon_i^2 = \sigma^2$ $(i = 1,\ldots, n)$, $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0$ $(i \neq j)$, но отказаться от нормальности распределения ε, то (7.9) также будет иметь место, но уже только для линейных несмещенных оценок $\tilde{\Theta}$.

В общем случае при нарушении (7.3) мнк-оценки теряют свои оптимальные свойства. Различные способы оценивания, применяемые в этом случае, описаны в § 7.2.

7.1.3. Ортогональная матрица плана. Матрицу $\mathbf{X}$ называют матрицей плана эксперимента. Рассмотрим случай, когда матрицу плана $\mathbf{X}$ можно разбить на k совокупностей столбцов $\mathbf{X}_1, \ldots, \mathbf{X}_k$ (что соответствует разбиению на k подмножеств анализируемого набора переменных) так, чтобы для всех $i \neq j$ столбцы матрицы $\mathbf{X}_i$ были ортогональны столбцам матрицы $\mathbf{X}_j$, т. е.

$$\mathbf{X}=(\mathbf{X}_1,\ \ldots,\ \mathbf{X}_k),\ \mathbf{X}_i'\,\mathbf{X}_j=0\ (i\neq j). \qquad (7.14)$$

Разобьем соответствующим образом и значения вектора

$$\Theta'=(\Theta^{(1)\prime},\ \ldots,\ \Theta^{(k)\prime}). \qquad (7.15)$$

Пусть далее r_i—ранг $\mathbf{X}_i$ и $\sum_{i=1}^{k} r_i = p$. Из (7.5) с учетом (7.14) и (7.15) получаем

$$\widehat{\Theta}=(\mathbf{X}'\,\mathbf{X})^{-1}\mathbf{X}'\,Y=\begin{bmatrix}\mathbf{X}_1'\,\mathbf{X}_1 & 0 & \ldots & 0\\ 0 & \mathbf{X}_2'\,\mathbf{X}_2 & \ldots & 0\\ \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \ldots & \mathbf{X}_k'\,\mathbf{X}_k\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1'Y\\ \mathbf{X}_2'Y\\ \cdot\ \cdot\\ \mathbf{X}_k'Y\end{bmatrix}=$$

$$=\begin{bmatrix}(\mathbf{X}_1'\,\mathbf{X}_1)^{-1} & \mathbf{X}_1'Y\\ (\mathbf{X}_2'\,\mathbf{X}_2)^{-1} & \mathbf{X}_2'\,Y\\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot\\ (\mathbf{X}_k'\,\mathbf{X}_k)^{-1} & \mathbf{X}_k'Y\end{bmatrix}=\begin{bmatrix}\widehat{\Theta}^{(1)}\\ \widehat{\Theta}^{(2)}\\ \ldots\\ \widehat{\Theta}^{(k)}\end{bmatrix}.$$

Другими словами, $\widehat{\Theta}^{(q)}$ является мнк-оценкой для $\Theta^{(q)}$ в модели $\mathrm{E}Y = \mathbf{X}_q\,\Theta^{(q)}$, а это означает, что $\Theta^{(q)}$ оцениваются независимо друг от друга и $\Theta^{(q)}$ не изменится, если положить какие-либо другие $\Theta^{(j)}$ $(j \neq q)$ равными нулю. Величина ОСК в рассматриваемом случае имеет вид:

$$\text{ОСК} = Y'Y - \widehat{\Theta}'\mathbf{X}'\mathbf{X}\widehat{\Theta} = Y'Y - \sum_{q=1}^{k} \widehat{\Theta}^{(q)}\mathbf{X}_q'\mathbf{X}_q\widehat{\Theta}^{(q)}. \qquad (7.16)$$

Если про какие-либо значения $\Theta^{(q)}$ известно, что они априори равны нулю, то соответствующие слагаемые в правой части (7.16) отсутствуют и ОСК соответственно больше. Поскольку $\Theta^{(q)}$ независимы между собой, то целесообразна независимая проверка гипотез $\Theta^{(q)} = \mathbf{0}$. Она проводится с помощью F-отношения:

$$F = \frac{\widehat{\Theta}^{(q)}\mathbf{X}_q'\mathbf{X}_q\widehat{\Theta}^{(q)}/r_q}{\text{ОСК}/(n-p)} = F(r_q, n-p). \qquad (7.17)$$

Это свойство широко используется в дисперсионном анали зе (см. гл. 13).

7.1.4. Параболическая регрессия и система ортогональных полиномов Чебышева. Пусть $\mathbf{E}y = \sum_{k=1}^{p} x^{k-1}\theta_k$. В силу соображений, изложенных в предыдущем пункте, целесообразно перейти к полиномам, ортогональным друг другу на системе наблюденных значений предиктора $x_1, \ldots, x_n$:

$$\psi_1(x) \equiv 1;$$

$$\psi_2(x) = x - \sum_{i=1}^{n} x_i/n = x - \frac{\Sigma x_i\psi_1(x_i)}{\Sigma\psi_1(x_i)\psi_1(x_i)} \cdot \psi_1(x);$$

$$\psi_3(x) = x^2 - \frac{\Sigma x_i^2\psi_2(x_i)}{\Sigma\psi_2(x_i)\psi_2(x_i)} \cdot \psi_2(x) - \frac{\Sigma x_i^2\psi_1(x_i)}{\Sigma\psi_1(x_i)\psi_1(x_i)} \cdot \psi_1(x);$$

. . . .

$$\psi_p(x) = x^{p-1} - \sum_{j=1}^{p-1} \frac{\Sigma x_i^{p-1}\psi_j(x_i)}{\Sigma\psi_j(x_i)\psi_j(x_i)} \cdot \psi_j(x) \quad (p > 3).$$

Введенные таким образом функции носят название ортогональных полиномов Чебышева. Соответствующие им столбцы матрицы плана $\mathbf{X}_k = (\psi_k(x_1), \ldots, \psi_k(x_n))'$, очевидно, ортогональны, и параметры в модели

$$\mathbf{E}y = \sum_{k=1}^{p} \psi_k(x)\theta_k \qquad (7.18)$$

оцениваются независимо друг от друга.

Когда истинный порядок полиномиальной регрессии не известен, то оценка параметров модели (7.18) проводится каж-

дый раз последовательно с проверкой гипотезы, что коэффициент перед очередным полиномом равен нулю. Как только эту гипотезу отвергнуть нельзя, подбор коэффициентов прекращается. Вопросы, связанные с последствием такого выбора правила остановки, обсуждались в гл. 6 (см. также [77, 147]).

7.1.5. Обобщенный мнк. Пусть теперь в модели (7.2)

$$\varepsilon \in N(0, \sigma^2 \mathbf{V}), \qquad (7.19)$$

где $\mathbf{V}$ — известная положительно определенная $(n \times n)$-матрица.

Важным примером подобной ситуации является случай, когда дисперсия ε зависит от значения регрессора X, но ε_i и ε_j при $i \neq j$ между собой некоррелированы. В этом случае в $\mathbf{V}$ отличны от нуля только диагональные элементы.

В общем случае отклики для различных значений предиктора, вообще говоря, зависимы. Но, что принципиально важно в постановке задачи, величина их корреляции известна априори.

Пусть $\mathbf{V} = \mathbf{C}\mathbf{C}'$ и $\tilde{Y} = \mathbf{C}^{-1}Y$, $\tilde{\mathbf{X}} = \mathbf{C}^{-1}\mathbf{X}$, $\varepsilon^* = \mathbf{C}^{-1}\varepsilon$. В новых переменных $\tilde{Y}$, $\tilde{\mathbf{X}}$, ε^* приходим к уравнению вида (7.2) $\tilde{Y} = \tilde{\mathbf{X}}\Theta + \varepsilon^*$, для которого (7.3) имеет место. Оценка Θ в преобразованной модели равна:

$$\Theta^* = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{Y} = (\mathbf{X}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{C}\mathbf{C}')^{-1}Y =$$
$$= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}Y. \qquad (7.20)$$

Очевидным образом модифицируются формулы (7.8) и (7.10):

$$s^{*2} = (Y - \mathbf{X}\Theta^*)'\mathbf{V}^{-1}(Y - \mathbf{X}\Theta^*); \qquad (7.21)$$

$$\Theta^* \in N(\Theta, \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}). \qquad (7.22)$$

7.2. Функции потерь, отличные от квадратичной

Мнк-оценки, получающиеся в результате минимизации выборочного критерия адекватности с квадратичной функцией потерь, неустойчивы к нарушениям предположения о нормальности распределения случайных ошибок. С утяжелением «хвостов» распределения они быстро теряют свои оптимальные свойства [14, п. 10.4.4]. Это связано с тем, что квадратичная функция потерь, используемая в мнк, придает слишком большой вес далеким отклонениям от регрессионной поверхности. Про-

гресс в области вычислительных методов позволяет перейти к использованию функций потерь $\rho(u)$, растущих при $|u| \to \infty$ более медленно, чем u^2. Соответствующие оценки по сравнению с мнк-оценками более устойчивы. Им и посвящен настоящий параграф. Так же, как при оценивании параметров положения и масштаба [14, п. 10.4.4], определенное внимание уделяется экспоненциально-взвешенным оценкам (эв-регрессии). Они допускают простую и наглядную интерпретацию, имеют хорошие выборочные свойства в случае небольших асимметричных искажений гауссовских распределений ошибок. Для них развита полная асимптотическая теория.

7.2.1. Функция потерь $\rho_\nu(u) = |u|^\nu$, $1 \leqslant \nu \leqslant 2$. Параметры регрессионной поверхности находят из условия минимизации по вектору Θ:

$$Q_\nu(\Theta) \equiv \sum_{i=1}^{n} \rho_\nu(u_i),$$

где $u_i = y_i - \sum_{k=1}^{p} x_i^{(k)} \theta_k$. Покажем, что для $\nu > 1$ 1) решение этой задачи $\widehat{\Theta}_\nu$ единственно;

2) в модели (7.2) для симметричных распределений случайных ошибок оценка $\widehat{\Theta}_\nu$ состоятельна. В самом деле, функция $\rho_\nu(u_i)$, рассматриваемая как функция от Θ, строго выпукла вниз. Следовательно, строго выпукла вниз и сумма $Q_\nu(\Theta)$, поэтому минимум $Q_\nu(\Theta)$ единствен и достигается в одной точке. Из строгой выпуклости $\rho_\nu(u)$ и, следовательно, положительности $\ddot{\rho}_\nu(u)$ вытекает, что для любой симметричной относительно нуля случайной величины ξ для любого $a \neq 0$

$$\mathbf{E}\rho_\nu(\xi + a) > \mathbf{E}\,\rho_\nu(\xi). \tag{7.23}$$

Из закона больших чисел [14, п. 7.2.1] следует, что в модели (7.2) для больших значений n для любого фиксированного вектора $C = (c^{(1)}, \ldots, c^{(p)})'$

$$Q_\nu(C)/n \approx \sum_{k=1}^{n} \mathbf{E}\rho_\nu(y_k - X_k' C)/n. \tag{7.24}$$

При симметричном относительно нуля распределении случайных ошибок, как следует из (7.23), правая часть (7.24) будет наименьшей при $C = \Theta$. Следовательно, в силу (7.24) $\widetilde{\Theta}_\nu$ должно быть при большом n близко к Θ, т. е. оценка $\widehat{\Theta}_\nu$ состоятельная.

В сформулированных выше условиях асимптотическая ковариационная матрица $\widehat{\Theta}_\nu$ имеет вид (см. также гл. 11):

$$\mathbf{E}(\widehat{\Theta}_\nu-\Theta)(\widehat{\Theta}_\nu-\Theta)'=\mathbf{K}^{-1}\frac{\mathbf{E}\dot{\rho}_\nu^2(\varepsilon)}{(\mathbf{E}\ddot{\rho}_\nu(\varepsilon))^2}, \tag{7.25}$$

где $\mathbf{K}=\sum_{i=1}^{n}X_iX'_i$; ε — случайный регрессионный остаток. В практической работе математические ожидания, стоящие в правой части (7.25), заменяются на их выборочные оценки:

$$\mathbf{E}\dot{\rho}_\nu(\varepsilon)\approx n^{-1}\sum_{i=1}^{n}\dot{\rho}_\nu^2(y_i-X'_i\hat{\Theta}_\nu); \tag{7.26}$$

$$\mathbf{E}\ddot{\rho}_\nu(\varepsilon)\approx n^{-1}\sum_{i=1}^{n}\ddot{\rho}_\nu(y_i-X'_i\hat{\Theta}_\nu). \tag{7.27}$$

Напомним, что формула (7.25) верна только для независимых и симметрично (относительно нуля) распределенных регрессионных остатков.

Методы вычисления $\widehat{\Theta}_\nu$ **[44, 94, 186]**. Основные уравнения имеют вид

$$\partial Q_\nu/\partial\theta_k=-\nu\sum_{i=1}^{n}\operatorname{sign}(u_i)\,|u_i|^{\nu-1}x_i^{(k)}=0,\ k=\overline{1,\ p}. \tag{7.28}$$

Введем под знак суммы веса $w_i=|u_i|^{\nu-2}$ и заменим $\operatorname{sign}(u_i)\ |u_i|^{\nu-1}$ на u_iw_i. Получим систему

$$\sum_{i=1}^{n}\left(y_i\,x_i^{(k)}-\sum_{j=1}^{p}x_i^{(j)}\,x_i^{(k)}\,\theta_j\right)w_i=0,\ k=\overline{1,\ p}. \tag{7.28'}$$

Система (7.28′) решается итеративно, при этом веса оцениваются на основе параметров, полученных на предыдущем шаге. В качестве нулевого приближения параметров можно взять обычные мнк-оценки. Чтобы не иметь дела со слишком большими весами, выбирают какую-либо большую константу $c>0$ и для $w_i\geqslant c$ полагают $w_i=c$. Для минимизации $Q_1(\Theta)$ пользуются также методами линейного программирования [253, 256] или специальным геометрическим приемом [53].

В качестве математической модели симметричного распределения с более тяжелыми хвостами, чем у нормального распределения, часто берут распределение Лапласа с плотностью

$$f(x)=\frac{1}{2\sigma}\exp\{-|x|/\sigma\}.$$

Если в модели (7.2) «остатки» ε_i не зависят от X_i, независимы между собой, одинаково распределены и имеют распределение Лапласа, то $\widehat{\Theta}_1$ есть оценка максимального правдоподобия для Θ

7.2.2. Оценка Хубера [213, 214]. Исходя из задачи поиска минимума максимальной (по всем симметричным засорениям нормального распределения) асимптотической дисперсии оценки параметра положения, П. Хубер ввел в рассмотрение функцию потерь

$$\rho_H(u) = \begin{cases} u^2/2, \text{ если } |u| < k; \\ k|u| - k^2/2, \text{ если } |u| \geqslant k. \end{cases} \tag{7.29}$$

Эта функция, являясь выпуклой, удачно сочетает достоинства ρ_2 при малых и умеренных значениях $|u|$ и ρ_1 — при больших отклонениях. Применение ρ_H для оценки регрессии в модели (7.2) требует обязательной одновременной оценки Θ и параметра масштаба распределения ε. Тем самым теряется одно из преимуществ ρ_υ — независимость процедур оценивания этих параметров. П. Хубер предложил искать $\widehat{\Theta}_H$ и $\widehat{\sigma}_H$ из решения системы:

$$\sum_{i=1}^{n} \dot{\rho}_H((y_i - X_i'\Theta)/\sigma) \cdot x_i^{(k)} = 0, \; k = \overline{1, p};$$
$$\sum_{i=1}^{n} \dot{\rho}_H^2((y_i - X_i'\Theta)/\sigma) = \beta, \tag{7.30}$$

где $\dot{\rho}_H(u) = d\rho_H(u)/du = \max(\min(k, u), -k)$, $\beta = (n-p)\int_{-\infty}^{\infty} \dot{\rho}_H^2(u)\,\varphi(u)\,du$ и $\varphi(u)$ — плотность стандартного нормального распределения.

Авторы [124] советуют заменить последнее уравнение в (7.30) на

$$\sigma^2 = \text{медиана } \{(y_i - X_i'\Theta)^2/a^2, \; i = 1, \ldots, n\}, \tag{7.31}$$

где $a \approx 0{,}675$. Теоретические свойства этих оценок и соответствующие вычислительные процедуры изучаются в [43, 110, 124]. На практике для оценки ковариационной матрицы $\widehat{\Theta}_H$ в случаях, когда распределение ε можно считать симметричным, можно использовать формулы (7.25), (7.26), (7.27) с заменой ρ_υ на ρ_H.

Оценки $\widehat{\Theta}_H$ все же еще недостаточно устойчивы к асимметричным отклонениям от нормальности распределения ε [149]. Следовательно, нужны функции потерь ρ, которые растут при $|u| \to \infty$ медленнее, чем ρ_1.

7.2.3. Функции потерь, имеющие горизонтальную асимптоту. Предложены три семейства функций, специально рассчитанных на асимметричные отклонения функции распределения ошибок от нормального закона. В унифицированных обозначениях в условиях, когда дисперсия основной (незасоренной) части распределения регрессионных остатков известна и равна единице, они могут быть приведены к виду (ниже параметр $\lambda > 0$):

функция потерь Андрюса [156]:

$$\rho_A(u) = \begin{cases} (2\lambda)^{-1}(1 - \cos((2\lambda)^{1/2}u)), & |u| < \pi(2\lambda)^{-1/2}; \\ \lambda^{-1}, & |u| \geqslant \pi(2\lambda)^{-1/2}; \end{cases} \quad (7.32)$$

функция потерь Мешалкина [89]:

$$\rho_M(u) = \lambda^{-1}(1 - \exp\{-\lambda u^2/2\}); \quad (7.33)$$

функция потерь Рамсея [243]:

$$\rho_R(u) = \lambda^{-1}(1 - (1 + \lambda^{1/2}|u|))\exp\{-\lambda^{1/2}|u|\}. \quad (7.34)$$

Все три функции при $\lambda \to 0$ стремятся к $u^2/2$, т. е. переходят в обычную квадратичную функцию потерь, используемую в мнк. При $\lambda \neq 0$ они имеют горизонтальную асимптоту, равную λ^{-1}. Взаимное расположение этих функций для двух значений параметров показано на рис. 7.1.

Так же, как при использовании ρ_H, в практической работе с этими функциями приходится выбирать значение параметра λ (настраивать ρ_A, ρ_M, ρ_R на определенный уровень отклонения от нормальности) и одновременно оценивать Θ и σ. При этом возникают дополнительные по сравнению с ρ_H трудности интерпретационного плана, связанные как с отсутствием выпуклости у новых функций, так и с сильным подавлением больших отклонений. Для иллюстрации сказанного рассмотрим пример. Пусть случайная величина ε дискретна, принимает всего два значения и $P\{\varepsilon = \pm \frac{3}{4}\pi(2\lambda)^{-1/2}\} = 0{,}5$. Несмотря на симметричность распределения ε относительно нуля $\mathbf{E}\,\rho_A(\varepsilon) > \min\limits_b \mathbf{E}\,\rho_A(\varepsilon - b)$, причем минимум достигается по крайней мере в двух различных точках. Это обстоятельство связано с локальной вогнутостью $\rho_A(u)$ в окрестности возможных значений ε. Аналогичное утверждение имеет место и для ρ_M и ρ_R.

В случае когда распределение ε (в общем случае не обязательно симметричное) сравнительно мало отличается от нормального закона N (...; 1), для всех трех функций ρ_A, ρ_M, ρ_R $a = \arg\min_b \mathbf{E}\rho(\varepsilon - b)$, т. е. значение b, при котором достигается минимум $\mathbf{E}\rho(\varepsilon - b)$, единственно и мало отличается от центра соответствующего нормального закона. Но оно, конечно, зависит от выбора функции ρ и от значения λ. Поэтому вопрос о содержательной интерпретации a остается акту-

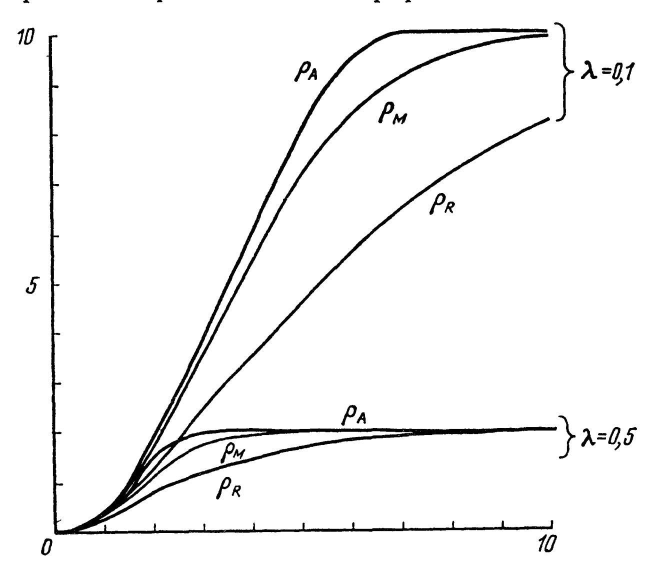


Рис. 7.1. Сравнение трех функций потерь при различных значениях λ

альным. В [14, п. 10.4.6) такая интерпретация описана для функции ρ_M (эв-оценки). По-видимому, аналогичная теория может быть построена и для ρ_A и ρ_R, но ρ_M несколько удобнее в аналитическом отношении, особенно в многомерном случае.

Вместе с тем вопрос, насколько распределение ε должно быть близко к нормальному закону для того, чтобы существовало только одно значение a, при котором достигается минимум $\mathbf{E}\rho(\varepsilon - b)$, пока не исследован достаточно подробно.

Было бы интересно сравнить оценки параметров, получаемые с помощью ρ_A, ρ_M и ρ_R, между собой в различных ситуациях. Но для этого требуется дальнейшее развитие теории устойчивого оценивания. Дело в том, что модель независимой выборки растущего объема из фиксированного распределения F_q, использованная Хубером, в которой $F_q(x-\mu) = (1-q)\times$

$\times \Phi (x - \mu) + qH (x - \mu)$, где Φ — функция нормального распределения, а H — функция распределения произвольного симметричного относительно нуля закона не очень подходит как из-за симметрии H, так и из-за того, что асимптотика, в которой q и H фиксированы, а объемы выборки $n \to \infty$, не вполне адекватна статистической практике: с ростом объема выборки мы узнаем F_q с возрастающей точностью и в принципе могли бы путем преобразования переменных усилить близость распределения к нормальному закону. Более адекватной моделью засорения является схема последовательности серий выборок растущего объема, в которой пропорция засорения $q = \gamma n^{-1/2}$ убывает с ростом n [149, 215 и 14, п. 6.1.11].

7.2.4. Эв-регрессия (λ-регрессия). Ниже, используя тот же методический прием,что и при введении эв-оценок [14, п. 10.4.6], с помощью цепочки определений вводится эв-регрессия и специальная мера отклонения от нее. Далее показывается, что эв-регрессия обладает рядом свойств, похожих на свойства обычной мнк-регрессии. Это облегчает содержательную интерпретацию эв-регрессии и выбор подходящего для конкретного случая значения λ. В заключение приводится асимптотическое разложение для оценок параметров эв-регрессии.

Пусть $w (y|X)$ — весовая функция y при фиксированном значении X, $F(...)$ — символ функции распределения. Введем

$$\mathrm{d}_w (X, F) = \int y w (y \mid X) \mathrm{d}F (y \mid X)/l_w (X, F); \quad (7.35)$$

$$g_w (X, F) = \int (y - \mathrm{d}_w (X, F))^2 w (y \mid X) \mathrm{d}F (y \mid X)/l (X, F), \quad (7.36)$$

где $l (X, F) = \int w (y \mid X) \mathrm{d}F (y \mid X)$.

Определение 7.1. Назовем $\mathrm{d}_w (X, F)$ — *w-взвешенной регрессией y* на X (w-взвешенным откликом в X), а $g_w(X, F)$ — *w-взвешенной дисперсией* относительно поверхности w-взвешенной регрессии.

Определение 7.2 Распределения $F (y \mid X)$ и $G (y \mid X)$ назовем *регрессионно-подобными*, если
$\mathrm{d}_w (X, F) = \mathrm{d}_w (X, G)$ и $g_w (X, F) = g_w (X, G)$.

Пусть $\varphi (y; a (X), \sigma^2 (X))$ — плотность нормального закона $N (y \mid X)$, (λ, c)-связанного с $F (y \mid X)$ [14, п. 10, 4.6], т. е. при $w (y \mid X) = \varphi^\lambda$ взвешенные моменты $N (y \mid X)$ и $F (y \mid X)$ совпадают.

Определение 7.3. Назовем $\Phi_\lambda (y, X)$ *λ-регрессионно-связанной* с $F (y, X)$, если
$\mathrm{d}\Phi_\lambda (y, X) = \varphi (y; a (X), \sigma^2 (X)) \mathrm{d}y \cdot \mathrm{d} F (X)$.

Определение 7.4. Назовем $a(X)$ — λ-регрессией (эв-регрессией) y на X и $\sigma^2(X)$ — *λ-дисперсией* y относительно поверхности λ-регрессии.

Аналог мнк утверждения для эв-регрессии. Пусть $\mathbf{E}_\lambda(y - c(X))^2 = \int\int (y - c(X))^2 \, d\Phi_\lambda(y, X)$, тогда $a(X) = \arg\min_{c(X)} \mathbf{E}_\lambda (y - c(X))^2$, т. е. если при каждом X заменить распределение $F(y, X)$ на λ-связанный с ним нормальный закон, то для нового распределения $\Phi_\lambda(y, X)$ $a(X)$ — обычная мнк-оценка регрессии y на X.

Пусть расстояние между двумя функциями распределения $F(y)$ и $G(y)$ определено как $\rho(F, G) = \sup_{a<b} \left| \int_a^b d(F - G) \right|$ и $\mathbf{M}(\varepsilon)$, ε — окрестность одномерных нормальных распределений, тогда для любого $\lambda > 0$ существуют такие $c = c(\lambda) > 0$ и $\varepsilon = \varepsilon(\lambda, c) > 0$, что для любого $F(y|X)$, для которого для всех X $F(y \mid X) \in \mathbf{M}(\varepsilon)$ существует единственная λ-регрессия y на X, причем $a(X)$ и $g(X)$ — непрерывные (в смысле ρ) функции относительно $F(y \mid X)$. Если $F(y \mid X)$ нормальны, то λ-регрессия y на X совпадает с обычной регрессией.

Таким образом эв-регрессия обладает всеми основными свойствами мнк-регрессии, только наблюдения в соответствующие формулы входят со специально подобранными весами. Введение весов позволяет как бы настраивать регрессию на интересующую исследователя часть выборки (рис. 7.2: в пунктирный овал заключены наблюдения (x_i, y_i), получившие малые веса и практически не участвующие в оценке параметров эв-регрессии; куполообразные кривые на прямой эв-регрессии показывают веса, приписанные наблюдениям). Эв-регрессия значительно устойчивее мнк-регрессии и регрессии по Хуберу к появлению далеких отклонений от регрессионной поверхности. Однако она, естественно, не является универсальным методом оценки регрессии для всех случаев, когда нарушаются предположения (7.3), лежащие в основе мнк. Четких рекомендаций, как выбирать λ в конкретном случае, пока не выработано. Ясно только, что надо давать максимальный вес «основной» части выборки и наименьший — части, где могут лежать «загрязнения». Определенные соображения по выбору величины λ в некоторых модельных случаях приведены в п. 7.2.5.

Минимизационное определение эв-регрессии. Для того чтобы охватить случай неизвестного σ, несколько изменим определение функции потерь по сравнению с (7.33). Пусть

$$\rho_\lambda(u, \sigma) = -\sigma^{-\lambda/(1+\lambda)} \exp\{-\lambda u^2/2\sigma^2\}, \qquad (7.37)$$

λ-регрессия $a(X) = X'\Theta$, где Θ — вектор неизвестных параметров, и для всех X λ-дисперсия $\sigma^2(X) = \sigma^2$ и σ^2 также неизвестно. Пусть далее $\gamma(\Theta, \sigma) = \mathbf{E}\rho_\lambda(y - X'\Theta, \sigma)$, где $\mathbf{E}$ — символ математического ожидания по мере $dF(y, X)$; тогда $\widehat{\Theta}_\lambda$ и $\widehat{\sigma}_\lambda$ являются решением уравнений

$$d\gamma/d\Theta = 0 \text{ и } d\gamma/d\sigma = 0 \quad (7.38)$$

и на них достигается локальный минимум $\gamma(\Theta, \sigma)$ (вообще говоря, не единственный).

Рассмотрим итерационную процедуру получения решения (7.38):

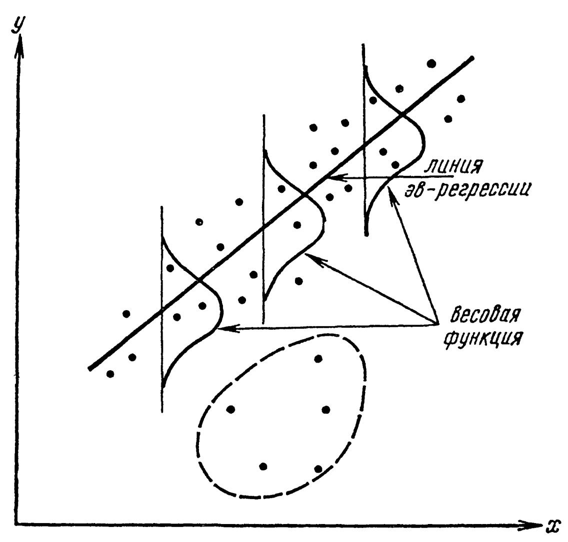


Рис. 7.2. Настройка эв-регрессии на интересующую исследователя часть выборки

$$\Theta_{(k+1)} = (\mathbf{E}XX'w_k)^{-1}\,\mathbf{E}yXw_k; \quad (7.39)$$

$$\sigma^2_{(k+1)} = (1+\lambda)\,\mathbf{E}(y - X'\Theta_{(k+1)})^2 w_k/\mathbf{E}w_k, \quad (7.40)$$

где $w_k = \exp\{-\lambda(y - X'\Theta)^2/2\sigma^2_{(k)}\}$.

Обозначим WREG оператор перехода по формулам (7.39), (7.40) от $(\Theta_{(k)}, \sigma^2_{(k)})$ к $(\Theta_{(k+1)}, \sigma^2_{(k+1)})$, тогда $(\Theta_\lambda, \sigma^2_\lambda)$ является неподвижной точкой оператора WREG.

Последнее определение λ-регрессии удобно для построения оценок $(\Theta_\lambda, \sigma^2_\lambda)$ по выборочным данным. Для этого достаточно в уравнениях (7.38) — (7.40) заменить символ математического ожидания $\mathbf{E}$ на знак суммирования по всем наблюдениям и ре-

шать их итерационно. Асимптотическая ковариационная матрица оценки $(\widehat{\Theta_\lambda, \sigma_\lambda^2})$, полученной по независимой выборке объема n, имеет вид

$$\mathbf{C} = n^{-1}\cdot \mathbf{K}^{-1}\, \mathbf{HK}, \tag{7.41}$$

где $\mathbf{K}$, $\mathbf{H}$ — квадратные матрицы порядка $(p+1)\times(p+1)$; $u = y - X'\Theta_\lambda$;

$$w = w(u, \sigma_\lambda) = \exp\{-\lambda u^2/2\sigma_\lambda^2\};\quad k_{ij} = \mathrm{E}x^{(i)}\, x^{(j)},\ i, j \leqslant p;$$

$$k_{i,\ p+1} = -\frac{\lambda(1+\lambda)}{2\sigma_\lambda^4}\cdot\frac{\mathrm{E}u^3\, w\cdot \mathrm{E}x^{(i)}}{\mathrm{E}w},\ i\leqslant p;$$

$$k_{p+1,\ p+1} = \frac{2+3\lambda}{4(1+\lambda)\sigma_\lambda^2} - \frac{\lambda(1+\lambda)}{4\sigma_\lambda^6}\cdot\frac{\mathrm{E}u^4 w}{\mathrm{E}w};$$

$$h_{i,j} = (1+\lambda)^2\frac{\mathrm{E}u^2 w^2}{(\mathrm{E}w)^2}\cdot \mathrm{E}x^{(i)}\, x^{(j)},\ 1\leqslant i, j\leqslant p;$$

$$h_{i,\ p+1} = \frac{1+\lambda}{2}\cdot\left[\frac{\mathrm{E}u^3 w^2}{\mathrm{E}uw\cdot \mathrm{E}w} - \frac{\mathrm{E}uw^2}{(\mathrm{E}w)^2}\right]\cdot \mathrm{E}x^{(i)},\ 1\leqslant i\leqslant p; \tag{7.42}$$

$$h_{p+1,\ p+1} = \frac{1}{4}\left[\frac{\mathrm{E}u^4 w^2}{(\mathrm{E}u^2 w)^2} + \frac{\mathrm{E}w^2}{(\mathrm{E}w)^2} - 2\,\frac{\mathrm{E}u^2 w^2}{\mathrm{E}u^2 w\cdot \mathrm{E}w}\right].$$

Для $\lambda > 0$ при любом $F(y \mid X)$ все входящие в формулы математические ожидания существуют. На практике их, а также σ_λ^2 и Θ_λ следует заменить соответствующими выборочными оценками.

7.2.5. Минимизация систематической ошибки. Практическое использование излагаемых выше предложений по повышению устойчивости оценок коэффициентов регрессии наталкивается на следующие неопределенности. Какую минимизируемую функцию риска выбрать? Все предлагаемые оценки содержат параметры: ν — в п. 7.2.1, k — в п. 7.2.2 и λ — в п. 7.2.3 и 7.2.4. Какими брать значения этих параметров? Если полезно уменьшать веса больших отклонений прогнозируемой переменной, то, может быть, полезно взвешивать и предикторные переменные?

В общем случае ответов на эти вопросы пока нет. Однако ориентиром может стать изучение модельных ситуаций. В частности, воспользуемся моделью засорения Шурыгина [14, п. 6.1.11]. В качестве основного распределения возьмем модель нормальной полиномиальной регрессии степени p, когда

плотность совместного распределения предикторной переменной x и прогнозируемой переменной y имеет вид

$$\varphi(x, y) = (2\pi\sigma_y \sigma_x)^{-1} \exp\left(-\frac{1}{2}\sigma_y^{-2}\left[y - \sum_{j=0}^{p} \theta_j x^j\right]^2 - \right.$$

$$\left. -\frac{1}{2}\sigma_x^{-2}[x-\mu]^2\right). \qquad (7.43)$$

Рассматривается серия q-засоренных выборок одинаковой длины n, и в k-й выборке засорение концентрируется в точке (x_k^*, y_k^*), так что плотности распределения выборок серии имеют вид

$$\begin{aligned} p_1(x, y) &= (1-q)\varphi(x, y) + q\delta(x_1^*, y_1^*); \\ p_2(x, y) &= (1-q)\varphi(x, y) + q\delta(x_2^*, y_2^*), \end{aligned} \qquad (7.44)$$

. . . .

где $\delta(x_k^*, y_k^*)$ — дельта-функция Дирака от точки (x_k^*, y_k^*). Пусть в серии выборок эта точка имеет плотность распределения $h(x^*, y^*)$.

Найдем квадратичную погрешность регрессионного предсказания $\widehat{y_0}$ для неизвестного значения результирующего показателя $y_0 = y(x_0)$, измеренного при $x = x_0$, когда двумерное распределение (x, y) описывается плотностью распределения (7.43), а прогноз $\widehat{y_0}$ строится по оценкам, основанным на произвольной выборке (7.44):

$$\mathbf{E}(\widehat{y_0} - y_0)^2 = \mathbf{E}\left(\sum_{j=0}^{p} \widehat{\theta}_j x_0^j - y_0\right)^2 =$$

$$= \mathbf{E}\left(\sum_{j=0}^{p} \theta_j x_0^j - y_0 + \sum_{j=0}^{p} \xi_i x_0^j\right)^2$$

(здесь усреднение производится и по x и по y, а $\xi_j = \widehat{\theta}_j - \theta_j$). Далее

$$\mathbf{E}(\widehat{y_0} - y_0)^2 = \mathbf{E}\left(\sum_{j=0}^{p} \theta_j x_0^j - y_0\right)^2 + 2\mathbf{E}\left(\sum_{j=0}^{p} \theta_j x_0^j - y_0\right) \times$$

$$\left(\sum_{j=0}^{p} \xi_j x_0^j\right) + \mathbf{E}\left(\sum_{j=0}^{p} \xi_j x_0^j\right)^2.$$

Первое слагаемое равно σ_y^2 и не может быть минимизировано. Во втором слагаемом сомножители независимы, и математическое ожидание первого из них равно нулю, так что нулю

равно все слагаемое. От способов оценивания коэффициентов регрессии зависит лишь *третье* слагаемое, которое, варьируя эти способы, можно минимизировать. Обозначив l-й момент стандартной нормальной величины через

$$\varkappa_l = \begin{cases} 1, \text{ если } l=0; \\ 0, \text{ если } l \text{ нечетно}; \\ 1\cdot 3\cdot 5 \ldots (l-1), \text{ если } l \text{ четно}, \end{cases}$$

третье слагаемое можно записать в виде

$$S^2 = \mathbf{E}\left(\sum_{j=0}^{p} \xi_j x_0^j\right)^2 = \sum_{j=0}^{p}\sum_{m=0}^{p} \varkappa_{j+m}\, \sigma_x^{j+m}\cdot \xi_j\, \xi_m.$$

Предположим для простоты, что величины σ_y^2, μ и σ_x^2 известны (устойчивые способы их оценки излагаются в [14, п. 10.4.4—10.4.6)]). Пусть оценки коэффициентов регрессии $\widehat{\theta_0}$, $\widehat{\theta_1}, \ldots, \widehat{\theta_p}$ находятся из системы уравнений

$$\sum_{i=1}^{n} K_j(x_i, y_i, \widehat{\theta_0}, \ldots, \widehat{\theta_p}) = 0, \; j=0, 1, \ldots, p$$

во всех выборках серии (7.44). Обозначим через $\mathbf{E}_k$ оператор математического ожидания, вычисляемого в соответствии с распределением $p_k(x,y)$ из (7.44). Мы можем искать минимум $\mathbf{E}_k S^2$ по способам оценивания коэффициентов регрессии.

Асимптотическое (при $n \to \infty$) поведение величины $\mathbf{E}_k S^2$ в любой из выборок (7.44) слагается из двух компонент: из дисперсии в модели (7.43) («случайная ошибка») и квадрата смещения за счет засорения («систематическая ошибка»). При росте n первая уменьшается как n^{-1}, вторая — как q^2. Между этими величинами возможны следующие соотношения:

а) величина q^2 уменьшается быстрее, чем n^{-1}. Тогда «систематическая ошибка» оказывается асимптотически пренебрежимой по сравнению со «случайной», имеющей порядок n^{-1}, классические оценки максимума правдоподобия оказываются асимптотически наилучшими, и приведенные выше рассуждения окажутся ненужными;

б) величина q^2 уменьшается медленнее, чем n^{-1}, например как $n^{-1+\gamma}$, где $0<\gamma<1$. Тогда дисперсия пренебрежимо мала по сравнению с квадратом смещения, и классические оценки не оптимальны, оптимальными будут оценки, минимизирующие «систематическую ошибку»; квадратическая погрешность оценки уменьшается не как n^{-1}, а медленнее, как $n^{-1+\gamma}$;

в) величины q^2 и n^{-1} имеют одинаковый порядок малости.

Этот вариант сводится к б) при рассмотрении *иерархии серий* [149].

Главный член асимптотического разложения $\mathrm{E}_h S^2$ в асимптотике б) определяется «систематической ошибкой» из-за засорения выборки и в среднем по серии (7.44), которое будем обозначать через $\mathrm{E}_{\text{сер}}$, зависит от плотности $h(x^*, y^*)$, так что существует

$$\omega^2 = \mathrm{E}_{\text{сер}} \lim_{n\to\infty} n^{1-\nu}\, \mathrm{E}_h S^2.$$

Используя известные методы минимаксной оптимизации, мы можем найти наилучшие оценки для наихудшей h, т. е. найти

$$\arg \min_{K_0, \ldots, K_p} \max_{h\in H} \omega^2. \qquad (7.45)$$

Результат зависит от множества H, среди которого отыскивается наихудшая h. Наиболее просто предположить, что h отличается от φ лишь значениями параметров: $h(x^*, y^*) = \varphi(x^*, y^*; \theta_0^*, \ldots, \theta_p^*, \sigma_y^{*2}, \mu^*, \sigma_x^{*2})$. В этом случае решение минимаксной задачи (7.45) приводит к следующей системе уравнений:

$$\hat\theta_0 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\hat\theta_j x_i^j\right) w_i \Big/ \sum_{i=1}^{n} w_i;$$

$$\hat\theta_1 = \sum_{i=1}^{n}\left(y_i - \hat\theta_0 - \sum_{j=2}^{p}\hat\theta_j x_i^j\right) x_i w_i \Big/ \sum_{i=1}^{n} x_i^2 w_i; \qquad (7.46)$$

. . . .

$$\hat\theta_p = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p-1}\hat\theta_j x_i^j\right) x_i^p w_i \Big/ \sum_{i=1}^{n} x_i^{2p} w_i,$$

где весовые функции

$$w_i = \exp\left(-\frac{1}{4\sigma_y^2}\left[y_i - \sum_{j=0}^{p}\hat\theta_j x_i^j\right]^2\right)\exp\left(-\frac{\nu_p}{2\sigma_x^2}[x_i - \mu]^2\right) \qquad (7.47)$$

являются экспонентами. Коэффициенты ν_p растут при росте p, оставаясь меньше единицы: $\nu_1 = 0{,}365$, $\nu_2 = 0{,}82$, $\nu_3 = 0{,}90$, ... Учитывая некоторую условность рассматриваемой модели, можно использовать аппроксимацию $\nu_p \approx (p - 0{,}83)/(p - 0{,}54)$. Система (7.46) решается итерациями.

Рассмотрим теперь задачу *нормальной многомерной линейной регрессии*, когда p предикторных переменных, образующих

вектор $X = (x^{(1)}, \ldots, x^{(p)})' \in N(\mu, \mathbf{C})$, используются для предсказания скалярной величины $y \in N(\theta_0 + \sum_{j=1}^{p} \theta_j x^{(j)}, \sigma^2)$, так что плотность совместного распределения X и y имеет вид

$$\varphi(X, y) = (2\pi)^{-\frac{p+1}{2}} |\mathbf{C}|^{-\frac{1}{2}} \sigma^{-1} \exp\left(-\frac{1}{2\sigma^2}\left[y - \theta_0 - \sum_{j=1}^{p} \theta_j x^{(j)}\right]^2 - \frac{1}{2}(X-\mu)' \mathbf{C}^{-1}(X-\mu)\right). \quad (7.48)$$

Рассмотрение аналогичной (7.44) схемы q-загрязненных выборок

$$\begin{aligned} p_1(X, y) &= (1-q)\varphi(x, y) + q\delta(X_1^*, y_1^*); \\ p_2(X, y) &= (1-q)\varphi(x, y) + q\delta(X_2^*, y_2^*) \end{aligned} \quad (7.49)$$

. . . .

и вполне аналогичная оптимизация погрешности предсказания $y_0 = y(X_0)$ с помощью регрессии y по X при известных σ^2, μ, $\mathbf{C}$ приводят к следующей системе уравнений для оценки коэффициентов регрессии:

$$\hat{\theta}_0 = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} \hat{\theta}_j x_i^{(j)}\right) w_i \Big/ \sum_{i=1}^{n} w_i;$$

$$\hat{\theta}_1 = \sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \sum_{j=2}^{n} \hat{\theta}_j x_i^{(j)}\right) x_i^{(1)} w_i \Big/ \sum_{i=1}^{n} (x_i^{(1)})^2 w_i; \quad (7.50)$$

. . . .

$$\hat{\theta}_p = \sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \sum_{j=1}^{p-1} \hat{\theta}_j x_i^{(j)}\right) x_i^{(p)} w_i \Big/ \sum_{i=1}^{n} (x_i^{(p)})^2 w_i,$$

где весовая функция также экспоненциальна:

$$w_i = \exp\left(-\frac{1}{4\sigma^2}\left[y - \hat{\theta}_0 - \sum_{j=1}^{p} \hat{\theta}_j x_i^{(j)}\right]^2\right) \exp\left(-\frac{\nu_p^*}{2} \times (X-\mu)' \mathbf{C}^{-1}(X-\mu)\right), \quad (7.51)$$

но величины ν_p^* убывают с ростом p:

$$\nu_p^* = \frac{1}{2}\left(\sqrt{\frac{p+5}{p+1}} - 1\right). \quad (7.52)$$

Сравним полученные оценки коэффициентов регрессии с излагавшимися в предыдущих разделах. Если весовые функции w_i положить равными единице, то системы (7.46) и (7.50) дадут оценки максимального правдоподобия соответственно для плотностей (7.43) и (7.48). Каждая из весовых функций (7.47) и (7.51) распадается на два экспоненциальных множителя, первая экспонента одинакова у обеих функций. Если вторые экспоненты заменить единицами, то решения совпадут с изложенной в предыдущем пункте эв-регрессией при $\lambda = 1/2$. Вторые экспоненты определяют взвешивание по предикторным переменным.

7.3. Байесовское оценивание

Общая методология байесовского оценивания описана в [14, п. 8.6.6]. Она сводится к введению априорной плотности распределения параметров и последующему нахождению по формуле Байеса с учетом экспериментальных данных их апостериорной плотности распределения. Ключевым моментом в применении байесовского оценивания является первый шаг.

7.3.1. Введение априорной плотности распределения параметров. Для априорных распределений возможны три интерпретации:

1) как частотных распределений;

2) как стандартных рекомендаций, что следует полагать о распределении неизвестных параметров в ситуации неопределенности;

3) как субъективной меры того, что полагает конкретный индивидуум.

Подробное обсуждение достоинств и недостатков этих подходов в общем случае может быть найдено в [70]. Здесь же мы ограничимся частной задачей — их использованием при оценке параметров регрессии.

Частотный подход. Предположим, что одна и та же регрессионная задача решается повторно на близком материале. Например, для разных районов страны изучается связь между производительностью труда и рядом параметров, характеризующих условия производства Или в медицине на материале различных медицинских центров по одним и тем же признакам строятся прогностические формулы для оценки риска осложнений какого-либо заболевания и т. п. Тогда в пространстве параметров, используемых в регрессионном уравнении, возникает эмпирическое распределение точек — оценок параметров, соответствующих отдельным решениям задачи (районам стра-

ны, медицинским центрам). После сглаживания оно может использоваться в качестве априорного распределения параметров регрессии. Этот подход является бесспорным с теоретической и практической точек зрения, но, к сожалению, довольно редко применимым, так как каждый исследователь стремится привнести что-либо свое в обработку, в набор регрессоров, численные значения, приписываемые градациям качественных регрессоров. Простое повторение проведенных другими исследований мало популярно. Другое дело, если обработка данных, собранных в разных местах по единой программе, проводится централизованно. В этом случае использование байесовского подхода может существенно уменьшить разброс в оценках параметров для каждого из массивов данных за счет привлечения к оцениванию информации о распределении параметров в других массивах.

Введение априорного распределения в ситуации неопределенности. Стандартный подход здесь заключается в том, что элемент априорной вероятности распределения $(\theta_1,\ldots, \theta_p,\sigma)$ в модели (7.2) берется пропорциональным [60]

$$d\theta_1\ldots d\theta_p d\sigma/\sigma. \tag{7.53}$$

Иногда говорят, что плотность априорного распределения пропорциональна $1/\sigma$:

$$p\ (\theta_1,\ldots,\ \theta_p,\ \sigma) \sim 1/\sigma. \tag{7.53'}$$

Правая часть (7.53′) не является плотностью в собственном смысле, так как интеграл от нее не определен, тем не менее при вычислении по формуле Байеса плотности апостериорного распределения параметров формальных трудностей при работе с (7.53) или не возникает, или они легко могут быть преодолены. Как мы увидим ниже в п. 7.3.2, выбор (7.53) удобен в аналитическом отношении и, казалось бы, хорошо отражает полное отсутствие априорных знаний о распределении параметров. Однако в нем на самом деле скрываются очень сильные предположения: отсутствие корреляции между параметрами (не путать с корреляцией между оценками значений параметров, которая зависит от распределения регрессоров и величины σ), пренебрежимая малость априорной вероятности того, что вектор параметров лежит в любом наперед заданном конечном объеме, какова бы ни была его величина, и т. д. Это приводит порою к серьезным трудностям с интерпретацией результатов байесовского оценивания [70].

Субъективный подход. В этом случае исследователь исходя из профессиональных соображений просто постулирует ап-

риорное распределение (Θ, σ). Для дальнейших расчетов удобны две формулы для априорной плотности. В первой из них распределение Θ не зависит от распределения σ:

$$p(\theta_1, \ldots, \theta_p, \sigma) = p_\Theta(\theta_1, \ldots, \theta_p) \cdot p_\sigma(\sigma) \sim |\mathbf{A}|^{1/2} \exp\left\{-(\Theta - \right.$$

$$\left. - \bar{\Theta})' \mathbf{A} (\Theta - \bar{\Theta})/2\right\} \sigma^{-(\nu_0+1)} \exp\left\{-\frac{\nu_0 c_0^2}{2\sigma^2}\right\}, \qquad (7.54)$$

где $\nu_0 > 0$, c_0, матрица $\mathbf{A}$ и вектор $\bar{\Theta}$ выбираются исследователем. При этом априорная ковариационная матрица компонент Θ $\mathbf{A}^{-1}$ предполагается невырожденной.

Во втором случае ковариационная матрица компонент пропорциональна σ^2 и

$$p(\theta_1, \ldots, \theta_p, \sigma) = \tilde{p}(\theta_1, \ldots, \theta_p | \sigma) p^*(\sigma) \sim \frac{|\mathbf{A}|^{1/2}}{\sigma^p} \times$$

$$\times \exp\{-(\Theta - \bar{\Theta})' \mathbf{A} (\Theta - \bar{\Theta})/2\sigma^2\} \cdot \sigma^{-(\nu_0+1)} \exp\left\{-\frac{\nu_0 c_0^2}{2\sigma^2}\right\}. \qquad (7.55)$$

Основная трудность субъективного подхода заключается в том, что информация, полученная из данных, рассматривается на равных основаниях с распределением, построенным исходя из не полностью формализованных соображений. Однако этот подход может быть полезен, когда выборка мала. Некоторые соображения в пользу (7.53′) приведены в п.7.3.3.

7.3.2. Апостериорное распределение параметров. В дальнейших расчетах предполагается, что имеют место базовые предположения мнк (7.2), (7.3), т. е.

$$p(Y | \mathbf{X}, \Theta, \sigma) \sim \frac{1}{\sigma^n} \exp\{-(Y - \mathbf{X}\Theta)'(Y - \mathbf{X}\Theta)/2\sigma^2\} =$$

$$= \frac{1}{\sigma^n} \exp\{-[(n-p)s^2 + (\Theta - \hat{\Theta})' \mathbf{X}' \mathbf{X} (\Theta - \hat{\Theta})]/2\sigma^2\},$$

где $s^2 = (Y - \mathbf{X}\hat{\Theta})'(Y - \mathbf{X}\Theta)/(n-p)$, $\hat{\Theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ — мнк-оценка Θ.

В предположении (7.53′),

$$p(\Theta, \sigma | Y, \mathbf{X}) \sim \frac{1}{\sigma} \cdot p(Y | \mathbf{X}, \Theta, \sigma). \qquad (7.56)$$

Откуда немедленно следует, что

$$p(\Theta \mid Y, \mathbf{X}) = \int\limits_0^{\infty} p(\Theta, \sigma \mid Y, \mathbf{X})\, d\sigma \sim [(n-p)s^2 +$$

$$+ (\Theta - \widehat{\Theta})' \mathbf{X}' \mathbf{X} (\Theta - \widehat{\Theta})]^{-n/2}, \qquad (7.57)$$

т. е. вектор Θ имеет так называемое многомерное распределение Стьюдента [60, с. 408—414]. Пусть m^{ii} — элемент матрицы $\mathbf{M}^{-1} = (\mathbf{X}'\mathbf{X})^{-1}$, тогда величина $(\theta_i - \widehat{\theta}_i)/s \cdot (m^{ii})^{1/2}$ имеет t-распределение Стьюдента с $(n-p)$ степенями свободы, что может быть использовано при построении одномерных доверительных интервалов для компонент Θ.

$$p(\sigma \mid Y, \mathbf{X}) = \int p(\Theta, \sigma \mid Y, \mathbf{X})\, d\Theta \sim \frac{1}{\sigma^{n-p+1}} \exp\{-(n-$$

$$-p)\, s^2/2\sigma^2\}, \qquad (7.58)$$

т. е. σ имеет обратное гамма-распределение, получаемое из обычного гамма-распределения [14, табл. 6.3] заменой аргумента x на $\sigma = 1/\sqrt{x}$.

Априорная плотность вида (7.55). В этом случае

$$p(\Theta;\ \sigma \mid Y, \mathbf{X}) \sim \frac{1}{\sigma^{n+p+\nu_0+1}} \exp\{-[\nu_0 c_0^2 + (\Theta - \overline{\Theta})' \times$$

$$\times \mathbf{A}(\Theta - \overline{\Theta}) + (Y - \mathbf{X}\Theta)'(Y - \mathbf{X}\Theta)]/2\sigma^2\} \sim$$

$$\sim \frac{1}{\sigma^{n'+p+1}} \exp\{-[n' c^2 + (\Theta - \breve{\Theta})' (\mathbf{A} + \mathbf{X}'\mathbf{X})(\Theta - \breve{\Theta})]/2\sigma^2\}, \qquad (7.59)$$

где $n' = n + \nu_0,\ n' c^2 = \nu_0 c_0^2 + Y'Y + \overline{\Theta}' \mathbf{A} \overline{\Theta} - \breve{\Theta}' (\mathbf{A} +$

$$+ \mathbf{X}'\mathbf{X}) \breve{\Theta} \text{ и } \breve{\Theta} = [\mathbf{A} + \mathbf{X}'\mathbf{X}]^{-1} (\mathbf{A}\overline{\Theta} + \mathbf{X}' Y). \qquad (7.60)$$

Интегрируя по σ, получаем апостериорную плотность

$$p(\Theta \mid Y, \mathbf{X}) \sim [n' c^2 + (\Theta - \breve{\Theta})' (\mathbf{A} + \mathbf{X}'\mathbf{X})(\Theta - \breve{\Theta})]^{-(n'+p)/2}. \qquad (7.61)$$

Априорная плотность вида (7.54). В этом случае, повторив с очевидными изменениями проведенные выше с плотностью вида (7.55) выкладки, получаем

$$p(\Theta \mid Y, \mathbf{X}) \sim \exp\{-(\Theta - \overline{\Theta})' \mathbf{A} (\Theta - \overline{\Theta})/2\} \cdot [\nu_0 c_0^2 +$$

$$+ Y'Y - \widehat{\Theta}' \mathbf{X}' \mathbf{X} \widehat{\Theta} + (\Theta - \widehat{\Theta})' \mathbf{X}\mathbf{X}' (\Theta - \widehat{\Theta})]^{-(n+\nu_0+p)/2}, \qquad (7.62)$$

где $\widehat{\Theta}$ — мнк-оценка Θ.

7.3.3. Повторная выборка из той же совокупности. Предположим, что из одной и той же совокупности делается повторная выборка, и обозначим Y_i, $\mathbf{X}_i$ вектор наблюдений и матрицу плана, относящиеся к i-й выборке ($i = 1,2$). Выбираем в качестве априорного распределения параметров для первой выборки (7.53′), тогда по (7.56) апостериорное распределение

$$p(\Theta, \sigma \mid Y_1, \mathbf{X}_1) \sim \frac{1}{\sigma^{n_1+1}} \exp\{-(Y_1-\mathbf{X}_1\Theta)'(Y_1- -\mathbf{X}_1\Theta)/2\sigma^2\} \sim \frac{1}{\sigma^{n_1+1}} \exp\{-[\nu_1 s_1^2 + (\Theta-\widehat{\Theta}_1)'\mathbf{X}_1'\mathbf{X}_1 \times \times (\Theta-\widehat{\Theta})]/2\sigma^2\}, \tag{7.63}$$

где $\nu_1 = n_1 - p$, $\widehat{\Theta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' Y_1$,

$\nu_1 s_1^2 = (Y_1 - \mathbf{X}_1\widehat{\Theta}_1)'(Y_1 - \mathbf{X}\widehat{\Theta}_1)$.

Заметим, что (7.63) имеет вид (7.55) с $\nu_0 = \nu_1$, $\overline{\Theta} = \widehat{\Theta}_1$, $\mathbf{A} = \mathbf{X}_1'\mathbf{X}_1$, $c_0^2 = s_1^2$, т. е. (7.63) можно рассматривать в качестве апостериорного распределения, полученного по байесовскому методу для некоторой выборки при стандартном выборе (7.53′) априорного распределения. Возьмем теперь (7.63) в качестве априорного распределения для второй выборки, тогда

$$p(\Theta, \sigma \mid Y_1, \mathbf{X}_1, Y_2, \mathbf{X}_2) \sim \frac{1}{\sigma^{n_1+n_2+1}} \exp\{-[(n_1-p)s_1^2 + +(\Theta-\widehat{\Theta}_1)'\mathbf{X}_1'\mathbf{X}_1(\Theta-\widehat{\Theta}_1) + (Y_2-\mathbf{X}_2\Theta)'(Y_2- -\mathbf{X}_2\Theta)]/2\sigma^2\} \sim \frac{1}{\sigma^{n_1+n_2+1}} \exp\{-[\nu s^2 + (\Theta- -\widetilde{\Theta})'\mathbf{M}(\Theta-\widetilde{\Theta})]/2\sigma^2\}, \tag{7.64}$$

где $\mathbf{M} = \mathbf{X}_1'\mathbf{X}_1 + \mathbf{X}_2'\mathbf{X}_2$, $\widetilde{\Theta} = \mathbf{M}^{-1}(\mathbf{X}_1 Y_1 + + \mathbf{X}_2 Y_2)$, $\nu s^2 = (Y_1 - \mathbf{X}_1\widetilde{\Theta})'(Y_1-\mathbf{X}_1\widetilde{\Theta}) + (Y_2 - \mathbf{X}_2\widetilde{\Theta})' \times \times (Y_2 - \mathbf{X}_2\widetilde{\Theta})$, $\nu = n_1 + n_2 - p$,

но это тот же вид, что в (7.63) для объединенной выборки. Таким образом, два процесса дают одно и то же апостериорное распределение параметров: 1) объединение массивов двух выборок с построением апостериорного распределения с использованием предположения (7.53′) и 2) использование предположения (7.53′) в качестве априорного только для первой выборки и получившегося апостериорного распределения для первой выборки в качестве априорного для второй.

7.4. Многомерная регрессия

При изучении эконометрических моделей (см. гл. 14), описании результатов сложных химических реакций, измерениях с помощью дублирующих приборов приходится сталкиваться с ситуацией, когда для каждого заданного значения регрессора $X = (x^{(1)}, \ldots, x^{(p)})'$ наблюдается не одномерный, как в предыдущих параграфах этой главы, а векторный отклик $Y = (y^{(1)}, \ldots, y^{(l)})'$. Соответствующую математическую задачу называют *многомерной регрессией*, или, более точно, многооткликовой регрессией (multiresponce regression) (п. 7.4.1). По сравнению с мнк-методом обычной регрессии (§ 7.1) оценка параметров множественной регрессии в общем случае усложняется, так как приходится одновременно оценивать параметры регрессионной зависимости и ковариационную матрицу случайных ошибок (п. 7.4.2). По аналогии с § 7.2 для многомерной регрессии удается построить оценки параметров, устойчивые к отклонениям от предположения нормальности распределения случайных ошибок (п. 7.4.3). В заключение обсуждается задача использования понятия множественной регрессии для параметризации распределения многомерного вектора (п. 7.4.4.).

7.4.1. Случай известной ковариационной матрицы ошибок. Пусть дана последовательность наблюдений (X_i, Y_i), $i = 1, \ldots, n$, и при этом предполагается, что

$$Y_i = \mathbf{f}(X_i, \Theta) + \boldsymbol{\varepsilon}_i, \; i = 1, \ldots, n, \tag{7.65}$$

где $\mathbf{f}(X_i, \Theta) = \mathbf{E}(Y_i \mid X_i)$ — l-мерная векторная регрессионная функция от X, известная с точностью до значения неизвестного векторного параметра $\Theta = (\Theta^{(1)}, \ldots, \Theta^{(l)})'$.

Рассмотрим модель, линейную относительно Θ (см. (7.1)):

$$\mathbf{f}(X, \Theta) = \boldsymbol{\Psi}'(X) \cdot \Theta,$$

где $$\boldsymbol{\Psi}(X) = \begin{Vmatrix} \Psi^{(1)}(X) & 0 & \ldots & 0 \\ 0 & \Psi^{(2)}(X) & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & \Psi^{(l)}(X) \end{Vmatrix}$$ — известная

матрица — функция от X, а

$$\Psi^{(j)}(X) = (\psi_{1j}(X), \ldots, \psi_{p_j j}(X))' \text{ и}$$

$$\Theta^{(j)} = (\theta_1^{(j)}, \ldots, \theta_{p_j}^{(j)})'.$$

Векторы случайных ошибок $\boldsymbol{\varepsilon}_i = (\varepsilon_i^{(1)}, \varepsilon_i^{(2)}, \ldots, \varepsilon_i^{(l)})$ взаимно независимы и имеют невырожденное l-мерное нормальное распределение с $\mathbf{E}\boldsymbol{\varepsilon} = 0$ и $\mathbf{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \mathbf{V}(X_i)$. Для упрощения обозначений будем писать $\boldsymbol{\Psi}_i$ и $\mathbf{V}_i$ вместо $\boldsymbol{\Psi}(X_i)$ и $\mathbf{V}(X_i)$.

В случае когда $\mathbf{V}_i$, известны и $\mathbf{M}_n$ определенное (7.67), начиная с некоторого n, имеет полный ранг, наилучшая линейная оценка для Θ имеет вид [117]:

$$\widehat{\Theta}_n = \mathbf{M}_n^{-1} Z_n, \tag{7.66}$$

где

$$\mathbf{M}_n = n^{-1} \sum_{i=1}^{n} \Psi_i \mathbf{V}_i^{-1} \Psi_i'; \tag{7.67}$$

$$Z_n = n^{-1} \sum_{i=1}^{n} \Psi_i \mathbf{V}_i^{-1} Y_i. \tag{7.68}$$

Формулы (7.66) — (7.68) легко могут быть получены из (7.20), если рассмотреть n наблюдений l-мерного вектора как $l \cdot n$ наблюдений одномерных векторов с известной блочно-диагональной (с блоками $\mathbf{V}_i$ размера $l \times l$) ковариационной матрицей между ними.

В сделанных предположениях оценка (7.66) состоятельна, несмещена и нормально распределена. Ее ковариационная матрица равна:

$$\mathbf{E}(\widehat{\Theta}_n - \Theta)(\widehat{\Theta}_n - \Theta)' = n^{-1} \mathbf{M}_n^{-1}. \tag{7.69}$$

7.4.2. Случай неизвестной ковариационной матрицы ошибок, не зависящей от значения предикторной переменной ($\mathbf{V}(X_i) \equiv$ $\equiv \mathbf{V}$). По аналогии с (7.66) в рассматриваемом случае оценка $\widehat{\Theta}$ находится из решения уравнения

$$\mathbf{M}_n(\Theta) \cdot \Theta = Z_n(\Theta), \tag{7.70}$$

где

$$\mathbf{M}_n(\Theta) = n^{-1} \sum_{i=1}^{n} \Psi_i \mathbf{V}^{-1}(\Theta) \Psi_i'; \tag{7.71}$$

$$Z_n(\Theta) = n^{-1} \sum_{i=1}^{n} \Psi_i \cdot \mathbf{V}^{-1}(\Theta) \cdot Y_i; \tag{7.72}$$

$$\mathbf{V}(\Theta) = n^{-1} \sum_{i=1}^{n} (Y_i - \Psi_i' \cdot \Theta)(Y_i - \Psi_i' \Theta)'. \tag{7.73}$$

Решение (7.70) удобно искать с помощью итерационной (по t) процедуры вида $\widehat{\Theta}_{t+1} = \mathbf{M}_n(\widehat{\Theta}_t) Z_n(\widehat{\Theta}_t)$. При выполнении дополнительного требования, что матрица

$$\mathbf{M} = \mathbf{E}\Psi(X) \mathbf{V}^{-1} \Psi(X) \tag{7.74}$$

невырождена, в [137] показано, что в окрестности истинного значения Θ итерационная процедура сходится с вероятностью, стремящейся к 1 при $n \to \infty$. В общем случае уже нельзя гарантировать единственность решения (7.70), а можно лишь утверждать, что при $n \to \infty$ среди решений (7.70) можно выделить последовательность $\widehat{\Theta}_n$, сходящуюся к истинному значению Θ. Эта подпоследовательность асимптотически-нормальна с параметрами Θ и n^{-1} **M**.

7.4.3. Эв-оценки. Введенное в п. 7.2.4 понятие экспоненциально-взвешенной регрессии (λ-регрессии) допускает естественное обобщение на случай многомерной регрессии. При этом сохраняется геометрическая интерпретация эв-регрессии с очевидным перенесением на многомерный отлик определений 7.1—7.4. Приведем только основные расчетные формулы, взяв за основу итерационный процесс, описанный в предыдущем пункте, и модифицировав его согласно (7.39), (7.40):

$$w_{\lambda,i,t} = \exp\{-\lambda (Y_i - \Psi_i' \Theta_t)' \mathbf{V}_\lambda^{-1} (\Theta_t) (Y_i - \Psi_i' \Theta_t)/2\}; \quad (7.75)$$

$$Z_\lambda (\Theta_t) = \sum_{i=1}^{n} w_{\lambda,i,t} \Psi_i \mathbf{V}_\lambda^{-1} (\Theta_t) Y_i \Big/ \sum_{i=1}^{n} w_{\lambda,i,t}; \quad (7.76)$$

$$\mathbf{M}_\lambda (\Theta_t) = \sum_{i=1}^{n} w_{\lambda,i,t} \Psi_i \mathbf{V}_\lambda^{-1} (\Theta_t) \Psi_i' \Big/ \sum_{i=1}^{n} w_{\lambda,i,t}; \quad (7.77)$$

$$\Theta_{t+1} = \mathbf{M}_\lambda^{-1} (\Theta_t) Z_\lambda (\Theta_t); \quad (7.78)$$

$$\mathbf{V}_\lambda (\Theta_{t+1}) = (1+\lambda) \sum_{i=1}^{n} w_{\lambda,i,t} (Y_i - \Psi_i' \Theta_{t+1}) \times$$

$$\times (Y_i - \Psi_i' \Theta_{t+1})' \Big/ \sum_{i=1}^{n} w_{\lambda,i,t}. \quad (7.79)$$

7.4.4. Использование многомерной регрессии для параметризации многомерных распределений. Плотность $p(X)$ распределения p-мерного случайного вектора $X = (X^{(1)}; X^{(2)})' = (x^{(1)}, \ldots, x^{(s)}, x^{(s+1)}, \ldots, x^{(p)})'$ всегда может быть представлена в виде $p(X) = p_1 (X^{(1)}) p_2 (X^{(2)}|X^{(1)})$. В гауссовском случае, когда

$p(X) = \varphi(X; M, \Sigma)$, а $M = (M^{(1)}, M^{(2)})$

и $\Sigma = \begin{Vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{Vmatrix}$ — вектор средних значений и ковариационная матрица, разбитые в соответствии с разбиением вектора X,

$$p(X) = \varphi(X^{(1)};\ M^{(1)}, \Sigma_{11})\, \varphi(X^{(2)};\ M^{(2)} - \Sigma_{21} \Sigma_{11}^{-1} (X^{(1)} - M^{(1)}),\ \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}).$$

Замечательная особенность многомерного нормального распределения состоит в том, что ковариационная матрица условного распределения $X^{(2)}$ при фиксированном значении $X^{(1)}$ не зависит от $X^{(1)}$ [20]. В общем случае это не так, и описание условного распределения значительно сложнее.

Для описания многомерного распределения предлагается распределение части координат ($X^{(1)}$) аппроксимировать стандартной нормальной моделью или считать таким, как оно получилось в выборке, а распределение остальных координат ($X^{(2)}$) заменить на надлежащим образом подобранный $(p-s)$-мерный нормальный закон со средним, линейно зависящим от $X^{(1)}$, и ковариационной матрицей $\mathbf{V}$ условного распределения $X^{(2)}$ при фиксированном значении $X^{(1)}$, от $X^{(1)}$ не зависящей. Но это и есть модель линейной многомерной регрессии, в которой $X^{(1)}$ играет роль предикторной точки-наблюдений (X), $X^{(2)}$ — роль многомерного результирующего показателя (Y), $\mathbf{E}(X^{(2)}|X^{(1)})$ — многомерная регрессия $X^{(2)}$ на $X^{(1)}$, а $X^{(2)} - \mathbf{E}(X^{(2)}|X^{(1)})$ — регрессионные остатки с ковариационной матрицей $\mathbf{V}$.

Если в основу подбора параметров многомерной регрессии при описании распределения $X^{(2)}$ положить требование совпадения не обычных, а *взвешенных* моментов условного распределения $X^{(2)}$ при известном значении $X^{(1)}$, то при соответствующем выборе весовой функции можно прийти к использованию эв-регрессии.

7.5. Оценивание параметров при наличии погрешностей в предикторных переменных *(конфлюэнтный анализ)*

7.5.1. Основные типы задач конфлюэнтного анализа. При анализе функциональных связей между переменными (см. § В.5, зависимости по схеме D) можно выделить следующие два случая.

1. Имеются две группы переменных $\eta \in R^m$ и $X \in R^p$. Переменные из первой группы известны экспериментатору со значительно бо́льшей ошибкой, чем из второй. В этом случае целесообразно работать с зависимостями вида

$$\eta = \mathbf{f}(X;\ \Theta). \tag{7.80}$$

Функция $\mathbf{f}(X;\ \Theta)$ предполагается заданной, и отыскание истинной зависимости заключается в оценивании параметров

$\Theta \in R^k$. Переменные X могут трактоваться как *предикторы* (предсказатели): задаваясь каким-либо их конкретным значением, можно предсказать значения переменных η. Если переменные X в процессе эксперимента могут изменяться по усмотрению экспериментатора, то говорят о *контролируемых переменных*. Переменные η часто называют *откликами*.

2. Если все переменные, с которыми имеет дело экспериментатор, известны примерно с одинаковой точностью, то имеет смысл использовать зависимости, представленные в виде

$$M(X; \Theta) = 0. \qquad (7.81)$$

При этом по-прежнему иногда удобно разделять переменные на результирующие (η) и объясняющие (контролируемые X:

$$M(\eta, X; \Theta) = 0: \qquad (7.81')$$

Ниже основное внимание уделяется *регрессионным* моделям, связанным с представлением (7.80), и лишь в заключительной части рассмотрены модели, порождаемые (7.81').

Регрессионные модели, связанные с (7.80). Возможно несколько постановок регрессионных задач, в основе которых лежит зависимость (7.80). Перечислим наиболее характерные из них. Читатель без особого труда сможет построить и некоторые промежуточные или смешанные конструкции.

Классическая регрессия (см. § 5.1.). В результате эксперимента (наблюдения) оказываются доступными величины $Y_i = \eta_i + \varepsilon_i$ и X_i, где ε_i — случайные величины (погрешности наблюдения). Иными словами,

$$Y_i = \mathbf{f}(X_i; \Theta_0) + \varepsilon_i, \; i = 1,2,\ldots, n. \qquad (7.82)$$

Подчеркнем, что значения предикторных переменных (условий наблюдения) известны точно. Нижний индекс 0 здесь и далее в этом параграфе обозначает истинное значение помеченной им величины.

Погрешности при фиксации условий наблюдения (активные эксперименты). Во многих экспериментах i-е наблюдение проводится при условиях X_i, несколько отличных от желаемых X_{0i}: $X_i = X_{0i} + \varepsilon_{Xi}$, где ε_{Xi} — случайные величины (погрешности фиксации). Таким образом, экспериментатору доступны величины $Y_i = \eta_i + \varepsilon_i$ и X_{0i}, связанные между собой соотношением

$$Y_i = \mathbf{f}(X_{0i} + \varepsilon_{Xi}, \Theta_0) + \varepsilon_i, \; i = 1, 2, \ldots, n. \qquad (7.83)$$

Пассивные наблюдения. Нередко (например, в эконометрических, социологических исследованиях) возможно лишь наб-

людение за *одновременным* изменением переменных η и X. Если эти наблюдения проводятся с некоторыми случайными погрешностями, то для анализа становятся доступными величины $Y_i = \eta_i + \varepsilon_i$ и $X_i = X_{0i} + \varepsilon_{Xi}$, или, в несколько более подробной записи,

$$\begin{cases} Y_i = \mathbf{f}(X_{0i}, \Theta_0) + \varepsilon_i; \\ X_i = X_{0i} + \varepsilon_{Xi}. \end{cases} \qquad (7.84)$$

По-видимому, впервые достаточно четкое разделение моделей (7.83) и (7.84) было осуществлено в [167], см. также [65, гл. 29].

Регрессионные задачи (7.82)—(7.84) содержат много общего как в постановке, так и в методах анализа. Более того, мы сознательно ограничимся рассмотрением именно тех методов, которые базируются на методе наименьших квадратов, широко используемом для классических регрессионных задач. В то же время внимание читателя будет обращено и на некоторые принципиальные различия в методах анализа соответствующих регрессионных задач.

7.5.2. Модифицированный мнк для схемы активного эксперимента. Обратимся вначале к регрессионной задаче (7.83), которая наиболее близка к классическому случаю.

Пусть анализируется *единственный* результирующий показатель ($m = 1$) и: а) случайные величины ε_i и $\varepsilon_{Xi} = \gamma \cdot \mathbf{v}_i$, фигурирующие в (7.83), независимы в совокупности и

$$\mathbf{E}\varepsilon_i = 0,\ \mathbf{E}\varepsilon_i^2 = \sigma^2,\ \mathbf{E}\varepsilon_{Xi} = 0,\ \mathbf{E}(\mathbf{v}_i \cdot \mathbf{v}_i') = \mathbf{d};$$

$$\mathbf{E}(|v_i^{(l)} v_i^{(q)} v_i^{(r)}|) \leqslant c < \infty,\ i = 1, 2, \ldots, n;\ l, q, r = 1, 2, \ldots, p;$$

c — некоторая константа, $\mathbf{d}$ — матрица $p \times p$, $\mathbf{v}_i$ — стандартизованная (например, $d_{ii} = 1$) случайная величина; б) существуют равномерно ограниченные на множестве допустимых условий наблюдения X_{0i}, $i = 1, \ldots, n$, производные по X функции $\mathbf{f}(X; \Theta)$ (см. (7.83)) до третьей включительно.

В рамках предположений а) и б) имеют место соотношения:

$$\mathbf{E}y_i = \mathbf{E}[f(X_{0i} + \gamma\mathbf{v}_i;\ \Theta_0) + \varepsilon_i] = \tilde{f}(X_{0i};\ \Theta_0) + 0(\gamma^3);$$

$$\mathbf{D}y_i = \lambda^{-1}(X_{0i};\ \Theta_0) + 0(\gamma^3), \qquad (7.85)$$

где $\mathbf{D}(\ )$, как обычно, означает дисперсию соответствующей случайной величины,

$$\tilde{f}(X_0;\ \Theta) = f(X_0;\ \Theta) + \frac{\gamma^2}{2} \cdot \mathrm{Sp}\left(\mathbf{d} \cdot \frac{\partial^2 f(X;\ \Theta)}{\partial X \partial X'}\bigg|_{X = X_0}\right);$$

$$\lambda^{-1}(X_0;\ \Theta) = \sigma^2 + \gamma^2 \frac{\partial f(X;\ \Theta)}{\partial X'} \cdot \mathbf{d} \cdot \frac{\partial f(X;\ \Theta)}{\partial X}\bigg|_{X = X_0}.$$

Заметим, что y_i, $i = \overline{1, n}$ — независимые случайные величины.

Таким образом, регрессионная задача (7.83) с точностью до $0(\gamma^3)$ сводится к регрессионной задаче

$$y_i = \tilde{f}(X_{0i};\ \Theta_0) + \mu_i, \tag{7.86}$$

где $\mathbf{E}\mu_i = 0$, $\mathbf{E}\mu_i^2 = \lambda^{-1}(X_{0i};\ \Theta_0)$.

Отличие (7.86) от (7.82) заключается в том, что *дисперсия погрешности зависит от неизвестных параметров* Θ_0, σ^2, $\mathbf{d}$. Подобным задачам посвящена довольно обширная литература (см., например, [12, 86, 138].

Остановимся на простейших оценках (σ^2 и $\mathbf{d}$ известны), предложенных в [182, 138]. Они определяются как предельная точка следующей итерационной процедуры:

$$\widehat{\Theta}_n = \lim_{s\to\infty} \Theta_s; \tag{7.87}$$

$$\Theta_s = \arg\min_{\Theta\in\Omega} n^{-1} \sum_{i=1}^{n} \lambda(X_{0i}, \Theta_{s-1})\ [y_i - \tilde{f}(X_{0i}, \Theta)]^2,$$

или ее модификацией, близкой к методу Ньютона — Рафсона:

$$\Theta_s = \Theta_{s-1} + \alpha_s \cdot Z_n^{-1}(\Theta_{s-1}) \cdot W_n(\Theta_{s-1}), \tag{7.88}$$

где

$$Z_n(\Theta) = n^{-1} \sum_{i=1}^{n} \lambda(X_{0i}, \Theta) \cdot \dot{F}(X_{0i}, \Theta)\ \dot{F}'(X_{0i}, \Theta);$$

$$W_n(\Theta) = n^{-1} \sum_{i=1}^{n} \lambda(X_{0i};\ \Theta) \cdot [y_i - \tilde{f}(X_{0i};\ \Theta)]\ \dot{F}(X_{0i};\ \Theta);$$

$$\dot{F}(X_0, \Theta) = \left.\frac{\partial \tilde{f}(X;\ \Theta)}{\partial \Theta}\right|_{X=X_0}.$$

Множитель α_s выбирается так же, как и в обычной процедуре Ньютона — Рафсона. Во избежание усложнений теоретического плана предполагается, что $\Theta_s \in \Omega$ для любого s.

Если в дополнение к условию а) из п. 7.5.2 и к условиям, сформулированным в комментариях к (7.85), потребовать: в) последовательность

$$v_n^2(\Theta) = n^{-1} \sum_{i=1}^{n} \lambda(X_{0i};\ \Theta_0)\, [\tilde{f}(X_{0i};\ \Theta) - \tilde{f}(X_{0i}\ \Theta_0)]^2$$

сходится равномерно по $\Theta \in \Omega$, причем $\lim v_n^2(\Theta) = v^2(\Theta)$, и функция $v^2(\Theta)$ имеет *единственный* минимум при $\Theta = \Theta_0$;

г) при всех $\Theta \in \Omega$ существуют непрерывные по Θ производные $\frac{d\tilde{f}(X_{0i};\ \Theta)}{d\Theta}$ и $\frac{d^2\tilde{f}(X_{0i};\ \Theta)}{d\Theta\, d\Theta'}$, и последовательности

$$\left\{n^{-1} \sum_{i=1}^{n} \lambda(X_{0i};\ \Theta_0)\, \varphi(X_{0i};\ \Theta) \cdot \psi(X_{0i};\ \Theta)\right\},$$

где функции $\varphi(X_{0i};\ \Theta)$ и $\psi(X_{0i};\ \Theta)$ могут совпадать с любой из указанных выше производных, сходятся равномерно по $\Theta \in \Omega$;

д) матрица $Z(\Theta_0) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \lambda(X_{0i};\ \Theta_0) \times$

$\times \dot{F}(X_{0i};\ \Theta_0) \cdot \dot{F}'(X_{0i};\ \Theta_0)$ —

неособенная,
тогда:

1) $\lim_{n\to\infty} P_n = 1$, где P_n — вероятность того, что при выборке объема n процедура сходится;

2) оценка $\widehat{\Theta}_n$, определяемая (7.87), сильно состоятельная, причем если при данных X_0 (7.87) имеет несколько решений, то за $\widehat{\Theta}_n$ принимается любое из них;

3) оценка $\widehat{\Theta}_n$ асимптотически-нормальная, т. е.

$$\lim_{n\to\infty} P\{\sqrt{n}\,(\widehat{\Theta}_n - \Theta_0) < t\} = \Phi_k(t;\ 0,\ Z^{-1}(\Theta_0)),$$

причем $Z_n(\widehat{\Theta}_n)$ (см. (7.88)) является сильно состоятельной оценкой матрицы $\mathbf{Z}(\Theta_0)$.

Отметим, что

$$\widehat{\Theta}_n \neq \arg\min_{\Theta\in\Omega} n^{-1} \sum_{i=1}^{n} \lambda(X_{0i};\ \Theta)\,[y_i - \tilde{f}(X_{0i};\ \Theta)]^2.$$

Данная теорема говорит о свойствах оценок для задачи (7.86). Для исходной регрессионной задачи все утверждения верны лишь в рамках приближения (7.85).

Оценки (7.87) достаточно просты как с точки зрения их статистического анализа, так и с вычислительной точки зрения. Однако они не могут быть использованы при неизвестных σ^2 и $\mathbf{d}$. Небольшое усложнение оценок (7.87) позволяет преодолеть

эту трудность. Рассмотрим следующую вспомогательную регрессионную задачу:

$$y_i = g(X_{0i};\ \Theta_0^*) + \xi_i;$$
$$\mathbf{E}\xi_i = 0,\ \mathbf{E}\xi_i^2 = \lambda^{-1}(X_{0i};\ \Theta_0^*). \tag{7.89}$$

В отличие от (7.85) в (7.89) не предполагается какой-либо специальной структуры $\lambda\ (X_{0i};\ \Theta_0^*)$. Более того, функции $g\ (X_0;\ \Theta_0^*)$ и $\lambda\ (X_0;\ \Theta_0^*)$ могут зависеть от *разных* групп параметров, входящих в Θ_0^*. Чтобы избежать непринципиальных усложнений, будем предполагать, что случайные величины ξ_i распределены нормально (в исходной задаче следует предположить нормальность ε и ν); при этом в (7.85) остаточный член, впрочем, как и для любого другого симметричного распределения, будет равен $O\ (\gamma^4)$.

Оценки параметров $\widehat{\Theta}^*$ определяются следующим образом:

$$\widehat{\Theta}^* = \lim_{s\to\infty} \Theta_s^*; \tag{7.90}$$

$$\Theta_s^* = \arg\min_{\Theta^*\in\Omega} n^{-1} \sum_{i=1}^{n} \Big\{\lambda\,(X_{0i},\ \Theta_{s-1}^*)\,[y_i - g\,(X_{0i},\ \Theta^*)]^2 +$$
$$+ \frac{1}{2}\lambda^2\,(X_{0i},\ \Theta_{s-1}^*)\,[\lambda^{-1}(X_{0i};\ \Theta^*) - (y_i - g\,(X_{0i};\ \Theta_{s-1}^*))^2]^2\Big\}.$$

Свойства оценок (7.90) можно проанализировать примерно так же, как это делается в [182] с оценками (7.87).

Предположим, что функции $g\ (X_0;\ \Theta^*)$ и $\lambda^{-1}\ (X_0;\ \Theta^*)$ удовлетворяют условиям, аналогичным (г), а функция

$$v_n^2\,(\Theta^*) = n^{-1} \sum_{i=1}^{n} \Big\{\lambda\,(X_{0i};\ \Theta^*)\,[y_i - g\,(X_{0i};\ \Theta^*)]^2 +$$
$$+ \frac{1}{2}\lambda^2\,(X_{0i};\ \Theta_0^*)\,[\lambda^{-1}(X_{0i};\ \Theta_0^*) - (y_i - g\,(X_{0i};\ \Theta_0^*))^2]^2\Big\}$$

условию в). Введем матрицу

$$\mathbf{G}_n\,(\Theta^*) = n^{-1} \sum_{i=1}^{n} [\lambda\,(X_{0i};\ \Theta^*)\,p\,(X_{0i};\ \Theta^*)\,p'\,(X_{0i};\ \Theta^*) +$$
$$+ \frac{1}{2}\lambda^2\,(X_{0i};\ \Theta^*)\,q\,(X_{0i};\ \Theta^*)\,q'\,(X_{0i};\ \Theta^*)],$$

где

$$p\,(X_0;\ \Theta^*) = \frac{\partial g\,(X_0;\ \Theta^*)}{\partial \Theta^*} \text{ и } q\,(X_0;\ \Theta^*) = \frac{\partial \lambda^{-1}\,(X_0;\ \Theta^*)}{\partial \Theta^*},$$

и потребуем (ср. с в)), чтобы существовала матрица $\mathbf{G}(\Theta_0^*) = \lim_{n\to\infty} \mathbf{G}_n(\Theta_0^*)$ и чтобы она была невырождена.

В рамках сделанных предположений:

1) $\lim_{n\to\infty} P_n = 1$, где P_n — вероятность того, что итерационная процедура (7.90) сходится при выборке объема n;

2) оценка $\widehat{\Theta}_n^*$ — сильно состоятельная, причем если при данном n имеется несколько решений, то за $\widehat{\Theta}_n^*$ принимается любое из них;

3) оценка $\widehat{\Theta}_n^*$ асимптотически-нормальна, т. е.

$$\lim P\{\sqrt{n}\,(\widehat{\Theta}_n^*-\Theta_0^*)<t\}=\Phi_k(t;\ 0,\ G^{-1}(\Theta_0^*)),$$

причем матрица $G_n(\widehat{\Theta}_n^*)$ является сильно состоятельной оценкой матрицы $G(\widehat{\Theta}_0^*)$.

Выше предполагалась нормальность распределения случайных величин ξ_i. Результаты остаются в силе, если потребовать, чтобы ξ_i имели конечные четыре момента, и заменить всюду «агрегат» $\frac{1}{2}\lambda(X_{0i};\ \Theta^*)$ на $m_4^{-1}(X_{0i},\ \Theta^*)$, где

$$m_4(X_{0i};\ \Theta^*)=\mathbf{E}\,[(\xi_i^2-\sigma_i^2)^2].$$

Конечно, на практике знание четырех моментов весьма проблематично. Но первые два пункта останутся справедливыми и без такой замены, хотя выражение для асимптотического значения дисперсионной матрицы примет при этом несколько более сложный вид. Интересно отметить, что в тех случаях, когда процедура (7.90) сходится, т. е. $\widehat{\Theta}_n$ определено, то предложенная оценка совпадает с оценкой максимального правдоподобия. Сформулированные утверждения позволяют получить некоторые полезные результаты для исходной задачи.

Если параметры σ^2 и $\mathbf{d}$ известны (см. комментарии к (7.87)), то $p(X_0;\ \Theta) = \dot{F}(X_0;\ \Theta)$ и

$$q(X_0;\ \Theta) = 2\gamma^2\ \frac{\partial \dot{F}(X;\ \Theta)}{\partial X}\cdot\mathbf{d}\cdot\frac{\partial f(X;\ \Theta)}{\partial X}\Bigg|_{X=X_0}=2\gamma^2\,\tilde{q}(X_0;\ \Theta).$$

Асимптотическое значение ковариационной матрицы $\Sigma_{\widehat{\Theta}}$ определяется матрицей

$$\mathbf{G}(\Theta_0^*)=\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}[\lambda(X_{0i};\ \Theta_0^*)\cdot\dot{F}(X_{0i};\ \Theta_0^*)\,\dot{F}'(X_{0i};\ \Theta_0^*)+$$

$$+\ 2\gamma^4\lambda^2(X_{0i};\ \Theta_0^*)\cdot\tilde{q}(X_{0i};\ \Theta_0^*)\cdot\tilde{q}'(X_{0i};\ \Theta_0^*)].$$

С точностью до О (γ^4) данная матрица совпадает с матрицей $\mathbf{Z}(\Theta_0)$, т. е. в рамках используемого приближения исходной задачи усложненная итерационная процедура (7.90) не приводит к оценкам асимптотически лучшим, чем (7.87). При неизвестных $\mathbf{d}$ и σ^2 можно без труда построить матрицу $\mathbf{G}(\Theta_0; \sigma_0^2; \mathbf{d}_0)$, имея в виду, что

$$\Theta^* = \begin{pmatrix} \Theta \\ \sigma^2 \\ \mathbf{d} \end{pmatrix}, \text{ где } \mathbf{d}' = (d_{11}\, d_{12} \ldots d_{jl} \ldots d_{pp});$$

$$p(X_0; \Theta^*) = \begin{pmatrix} \dot{F}(X_0; \Theta) \\ 0 \\ \frac{\gamma^2}{2} \ddot{F}(X_0; \Theta) \end{pmatrix},$$

$$\text{где } \ddot{F}'(X_0; \Theta) = \left(\frac{\partial^2 f(X_0; \Theta)}{\partial x_0^{(1)^2}} \quad \frac{\partial^2 f(X_0; \Theta)}{\partial x_0^{(1)} \partial x_0^{(2)}} \cdots \frac{\partial^2 f(X_0; \Theta)}{\partial x_0^{(p)^2}} \right);$$

$$q(X_0; \Theta^*) = \begin{pmatrix} 2\gamma^2 \tilde{q}(X_0; \Theta) \\ 1 \\ \gamma^2 \dot{F}^{(2)}(X_0; \Theta) \end{pmatrix},$$

где

$$\dot{F}^{(2)'}(X_0; \Theta) = \left(\left(\frac{\partial f(X_0; \Theta)}{\partial x_0^{(1)}} \right)^2 \frac{\partial f(X_0; \Theta)}{\partial x_0^{(1)}} \cdot \frac{\partial f(X_0; \Theta)}{\partial x_0^{(2)}} \cdots \right.$$

$$\left. \cdots \left(\frac{\partial f(X_0; \Theta)}{\partial x_0^{(p)}} \right)^2 \right).$$

Из двух последних формул видно, что σ^2 и $\mathbf{d}$ оцениваемы раздельно, если компоненты вектора $(\mathbf{O}, \ddot{F}')$ или вектора $(1, \dot{F}^{(2)'})$ линейно-независимы на множестве точек $X_1, \ldots, X_n$.

7.5.3. Пассивные наблюдения. В теоретическом плане регрессионная задача, определяемая (7.84) и условиями а) из п.7.5.2, оказывается существенно сложнее регрессионной задачи (7.83). Тем не менее ввиду своей актуальности она уже давно привлекла внимание статистиков. По-видимому, первая работа, посвященная задаче (7.84), появилась в 1901 г. [235] и содержала идею, лежащую в основе практически всех результатов, связанных с упомянутой задачей. Идея предельно проста: за оценки параметров Θ принимать те значения, при которых минимально суммарное расстояние точек (X_{0i}, y_i) от поверхности $y = f(X; \Theta)$ в легко интерпретируемой метрике, т. е.

$$\hat{\Theta} = \arg \min_{\Theta \in \Omega} \sum_{i=1}^{n} l_i^2(\Theta), \tag{7.91}$$

где

$$l_i^2(\Theta)=\min_{X_i}\left[\left(\frac{y_i-f(X_i;\ \Theta)}{\sigma_i}\right)^2+\gamma^{-2}(X_{0i}-X_i)'\,\mathbf{d}^{-1}(X_{0i}-X_i)\right].$$

Оценки (7.91) называют *оценками метода наименьших расстояний.* Ниже рассмотрен приближенный вариант этих оценок, позволяющий обойтись численными процедурами, развитыми для метода наименьших квадратов. Несложные вычисления приводят в линейном случае ($f(X;\ \Theta)=\Theta'X$) к простой формуле $l_i^2=(y_i-\Theta'X_{0i})^2/(\sigma^2+\gamma^2\Theta'\,\mathbf{d}\Theta)$. В случае произвольной функции $f(X;\ \Theta)$ (но имеющей необходимое количество производных) и при ошибках ε_i и $\mathbf{v}_i$, удовлетворяющих условиям а) из п. 7.5.2, имеет место приближенная формула

$$l_i^2=(y_i-\bar{f}(X_{0i};\ \Theta))^2\lambda(X_{0i};\ \Theta)+0(\gamma^3), \tag{7.92}$$

где

$$\bar{f}(X_{0i};\ \Theta)=f(X_{0i};\ \Theta)-\frac{\gamma^2}{2}\operatorname{Sp}\left(\mathbf{d}\cdot\frac{\partial^2 f(X;\ \Theta)}{\partial X\partial X'}\bigg|_{X=X_0}\right);$$

$$\lambda^{-1}(X_0;\ \Theta)=\sigma^2+\gamma^2\frac{\partial f(X;\ \Theta)}{\partial X'}\cdot\mathbf{d}\cdot\frac{\partial f(X;\ \Theta)}{\partial X}\bigg|_{X=X_0}.$$

Определим оценки следующим образом:

$$\widehat{\Theta}_n=\arg\min_{\Theta\in\Omega}\sum_{i=1}^{n}\lambda(X_{0i};\ \Theta)(y_i-\bar{f}(X_{0i};\ \Theta))^2. \tag{7.93}$$

Введем функции

$$\bar{\dot{F}}(X;\ \Theta)=\frac{\partial\bar{f}(X;\ \Theta)}{\partial\Theta}\ \text{и}\ \bar{\mathbf{d}}(X;\ \Theta)=\frac{\partial\dot{F}(X;\ \Theta)}{\partial X'}\cdot\mathbf{d}\cdot\frac{\partial\dot{F}(X;\ \Theta)}{\partial X}.$$

Пусть выполняются условия а) — д) из п. 7.5.2 с очевидной заменой $\dot{F}(X_0;\ \Theta)$ на $\bar{\dot{F}}(X;\ \Theta)$ и $\tilde{f}(X_0;\ \Theta)$ на $\bar{f}(X;\ \Theta)$ и дополнительно существует предел

$$\mathbf{z}(\Theta_0)=\lim_{n\to\infty}n^{-1}\sum_{i=1}^{n}\lambda(X_i;\ \Theta_0)\cdot\bar{\mathbf{d}}(X_i;\ \Theta_0);$$

тогда в рамках приближения (7.92):

1) оценка (7.93) и сильно состоятельна и асимптотически-нормальна, т. е.

$$\lim_{n\to\infty}P\{\sqrt{n}\,(\widehat{\Theta}_n-\Theta_0)<t\}=\Phi_k(t;\ 0,\ \Sigma_{\widehat{\Theta}}),$$

где

$$\Sigma_{\widehat{\Theta}} = \mathbf{Z}^{-1}(\Theta_0)\,[\mathbf{Z}(\Theta_0) + \gamma^2 \mathbf{z}(\Theta_0)]\,\mathbf{Z}^{-1}(\Theta_0);$$

2) сильно состоятельными оценками матриц $\mathbf{Z}(\Theta_0)$ и $\mathbf{z}(\Theta_0)$ являются соответственно матрицы

$$n^{-1}\sum_{i=1}^{n} \lambda(X_{0i};\ \widehat{\Theta}_n)\,[\dot{\bar{F}}(X_{0i};\ \widehat{\Theta}_n)\,\dot{\bar{F}}'(X_{0i};\widehat{\Theta}_n) - \gamma^2 \bar{\mathbf{d}}(X_{0i};\widehat{\Theta}_n)];$$

$$n^{-1}\sum_{i=1}^{n} \lambda(X_{0i};\ \widehat{\Theta}_n)\cdot\bar{\mathbf{d}}(X_{0i};\widehat{\Theta}_n).$$

Ряд полезных результатов, описывающих поведение «приближенных» оценок в рамках исходной модели (7.91), обсуждается в [81].

При подсчете оценок (7.93) оказывается удобным введение фиктивных наблюдений $y_{ф} \equiv 0$ и отклика

$$\bar{\bar{f}}(X_{0i};\ \Theta) = \lambda^{1/2}(X_{0i};\ \Theta)\,[y_i - \bar{f}(X_{0i};\ \Theta)].$$

Для минимизации функции

$$v_n^2(\Theta) = n^{-1}\sum_{i=1}^{n} (y_{ф i} - \bar{\bar{f}}(X_{0i};\ \Theta))^2$$

можно обратиться к любой программе нелинейного мнк. Обычно в этих программах в качестве оценки ковариационной матрицы используется матрица $\widehat{\Sigma}_{\widehat{\Theta}}(n) = \bar{\bar{\mathbf{Z}}}^{-1}_n$,

где

$$\bar{\bar{\mathbf{Z}}}_n = n^{-1}\sum_{i=1}^{n} \frac{\partial \bar{\bar{f}}(X_{0i};\ \Theta)}{\partial \Theta}\cdot\frac{\partial \bar{\bar{f}}(X_{0i};\ \Theta)}{\partial \Theta'}\bigg|_{\Theta=\widehat{\Theta}_n}.$$

Можно показать, что

$$\bar{\bar{\mathbf{Z}}}_n \to \mathbf{Z}(\Theta_0) + \gamma^2 \bar{\bar{z}}(\Theta_0) \text{ при } n \to \infty \text{ почти наверное,}$$

т. е. $\widehat{\Sigma}_{\Theta}(n)$ является *заниженной* оценкой дисперсионно-ковариационной матрицы $\Sigma_{\widehat{\Theta}}$.

7.5.4. Некоторые принципиальные отличия регрессионных задач (7.83) и (7.84). Как нетрудно видеть, в первом случае любая оценка вида

$$\widetilde{\Theta}_n = \arg\min_{\Theta\in\Omega} n^{-1}\sum_{i=1}^{n} \omega_i\,[y_i - f(X_{0i};\ \Theta)]^2$$

является сильно состоятельной. Оценка (7.87) является по эффективности асимптотически эквивалентной наилучшей (т. е. с оптимально выбранными весами ω_i) среди них. Для задачи (7.83) оказывается возможным состоятельно оценить параметры Θ, не используя информации о дисперсиях σ^2 и $\mathbf{d}$. Этот факт позволяет в свою очередь говорить о состоятельном оценивании параметров σ^2 и $\mathbf{d}$ (см (7.90)). В случае (7.84) отсутствие информации о σ^2 и $\mathbf{d}$ не позволяет построить состоятельных оценок параметров Θ. Подобный результат был впервые отмечен, по-видимому, в [223].

Другим, менее существенным отличием является знак поправки к функции $f\ (X_0;\ \Theta)$ в схемах активного и пассивного экспериментов.

7.5.5. Неявное задание отклика. В тех случаях, когда переменные, подверженные ошибкам, не разделяются естественным образом на две группы (зависимые переменные и предикторы), целесообразно обратиться к рассмотрению неявных зависимостей (7.81).

Будем рассматривать модель

$$M\,(X_0;\ \Theta)=0,\ X_i=X_{0i}+\varepsilon_i,\ i=\overline{1,\ n}. \tag{7.94}$$

Экспериментатору известна функция $M\ (X_0;\ \Theta)$ и искаженные наблюдения X_i. Переменные η, входящие в (7.81′), опущены, так как их введение непринципиально для последующих результатов.

Предположим, что ошибки $\varepsilon_i=\gamma\mathbf{v}_i$ удовлетворяют следующим правилам: они независимы в совокупности и

$$\mathbf{E}\mathbf{v}_i=0,\ \mathbf{E}\,(\mathbf{v}_i\,\mathbf{v}_i')=\mathbf{d},\ \mathbf{E}\,(|\,v_i^{(l)}\,v_i^{(q)}\,v_i^{(r)}|)\leqslant c<\infty,$$

$$i=1,\ 2,\ \ldots,\ n;\ l,\ q,\ r=1,\ 2,\ \ldots,\ p.$$

Так же, как и выше, рассмотрим оценки метода наименьших расстояний:

$$\widehat{\Theta}_n=\arg\min_{\Theta\in\Omega}\sum_{i=1}^{n}l_i^2\,(\Theta), \tag{7.95}$$

где

$$l_i^2(\Theta)=\min_{X_{0i}}\,(X_i-X_{0i})'\,\mathbf{d}^{-1}\,(X_i-X_{0i});\ M\ (X_{0i};\ \Theta)=0.$$

Пусть вначале $M\,(X_0;\ \Theta)=1+\Theta'\,X_0$. Тогда

$$\widehat{\Theta}_n=\arg\min_{\Theta\in\Omega}\sum_{i=1}^{n}\frac{(1+\Theta'\,X_i)^2}{\Theta'\,\mathbf{d}\Theta},$$

что, по существу, совпадает с (7.91).

Если $M(X_0; \Theta) = 1 + \Theta' \psi(X_0)$, то, полагая

$$\widetilde{X}_{0i} = \psi(X_{0i}),$$

$$\widetilde{X}_i = \psi(X_i) - \frac{\gamma^2}{2} \operatorname{Sp}\left(\frac{\partial^2 \psi(X)}{\partial X \partial X'} \cdot \mathbf{d} \Big|_{X = X_i} \right),$$

$$\widetilde{\mathbf{d}}_i = \frac{\partial \psi(X)}{\partial X'} \cdot \mathbf{d} \cdot \frac{\partial \psi(X)}{\partial X} \Big|_{X = X_i},$$

с точностью до $O(\gamma^3)$ (7.85) можно переписать в виде

$$\widehat{\Theta}_n = \arg\min_{\Theta \in \Omega} \sum_{i=1}^{n} \frac{(1 + \Theta' \widetilde{X}_i)^2}{\Theta' \cdot \widetilde{\mathbf{d}}_i \cdot \Theta}.$$

В случае произвольной параметризации с точностью до $O(\gamma^3)$:

$$\widehat{\Theta}_n = \arg\min_{\Theta \in \Omega} \sum_{i=1}^{n} \frac{\widetilde{M}^2(X_i; \Theta)}{\psi(X_i; \Theta)},$$

где

$$\widetilde{M}(X_i; \Theta) = M(X_i; \Theta) - \frac{\gamma^2}{2} \operatorname{Sp}\left(\frac{\partial^2 M(X, \Theta)}{\partial X \partial X'} \Big|_{X = X_i} \right);$$

$$\psi(X_i; \Theta) = \frac{\partial M(X; \Theta)}{\partial X'} \cdot \mathbf{d} \cdot \frac{\partial M(X; \Theta)}{\partial X} \Big|_{X = X_i}.$$

Для оценок $\widehat{\Theta}_n$ имеют место результаты, практически полностью аналогичные изложенным в п. 7.5.3.

В заключение отметим, что в [86] можно найти описание конкретных реализаций на ЭВМ описанных выше алгоритмов.

7.6. Оценивание в регрессионных моделях со случайными параметрами *(регрессионные задачи второго рода)*

7.6.1. Постановка задачи. Рассмотрим следующую модель:

$$y_{ij} = \Theta_j' \mathbf{f}(X_{ij}) + \varepsilon_{ij}, \quad i = \overline{1, n_j}, \quad j = \overline{1, k}, \tag{7.96}$$

где случайные величины ε_{ij} при фиксированном j удовлетворяют стандартным требованиям: $\mathbf{E}\varepsilon_{ij} = 0$, $\mathbf{E}(\varepsilon_{ij} \cdot \varepsilon_{i'j}) = \delta_{ii'} \cdot \sigma^2$. Параметры Θ_j предполагаются случайными, причем $\Theta_j' =$

$= (\theta_{1j}, \ldots, \theta_{lj})$, $\mathbf{E}\Theta_j = \Theta_0$, $\mathbf{E}[(\Theta_j - \Theta_0)(\Theta_j - \Theta_0)'] = \Sigma$. В зависимости от постановки задачи вектор Θ_0 и матрица Σ могут быть или заданы, или неизвестны. Величины Θ_j и ε_{ij} предполагаются некоррелированными.

В практических исследованиях индексом j может служить, например, номер предприятия из совокупности аналогичных, номер партии сырья, номер пациента из группы больных, подвергающихся одному и тому же способу лечения, и т. д.

В тех случаях, когда выясняется поведение *каждого* j-го объекта, необходимо решать задачу об оценивании параметров Θ_j, $j = \overline{1,k}$. Если же необходимо понять поведение *всей* совокупности объектов, то приходится говорить об оценивании вектора Θ_0.

7.6.2. Случай, когда средние значения Θ_0 и ковариационная матрица Σ оцениваемых параметров известны (требуется оценить параметры Θ_j). Необходимо, чтобы оценки удовлетворяли следующим требованиям (ср. с. § 7.1):

$$\widehat{\Theta}_j = \mathbf{L}_1 \mathscr{Y}_j + \mathbf{L}_2 \Theta_0;$$

$$\mathbf{E}\widehat{\Theta}_j = 0; \qquad (7.97)$$

$$\mathbf{D}(\widehat{\Theta}_j) = \mathbf{E}[(\widehat{\Theta}_j - \Theta_j)(\widehat{\Theta}_j - \Theta_j)'] = \min_{L_1,\, L_2} \mathbf{E}(\mathbf{L}_1 \mathscr{Y}_j + \mathbf{L}_2 \Theta_0 - \Theta_j)(\mathbf{L}_1 \mathscr{Y}_j + \mathbf{L}_2 \Theta_0 - \Theta_j)'],$$

где

$$\mathscr{Y}_j = (y_{1j}\, y_{2j} \ldots y_{n_j j})'.$$

В (7.97) подразумевается, что матрица $\mathbf{B}^*$ является решением экстремальной задачи $\mathbf{A}(\mathbf{B}^*) = \min\limits_{\mathbf{B} \in \mathscr{B}} \mathbf{A}(\mathbf{B})$, где $\mathbf{A}(\mathbf{B})$ — положительно полуопределенная матрица при всех допустимых $\mathbf{B}$, если выполняется матричное неравенство $\mathbf{A}(\mathbf{B}) \geqslant \mathbf{A}(\mathbf{B}^*)$.

По аналогии с § 7.1 нетрудно получить (см. например, [135]), что

$$\widehat{\Theta}_j = (\mathbf{M}_j + \Sigma^{-1})^{-1}(\Sigma^{-1}\Theta_0 + Y_j); \qquad (7.98)$$

$$\mathbf{D}(\widehat{\Theta}_j) = (\mathbf{M}_j + \Sigma^{-1})^{-1},$$

где

$$\mathbf{M}_j = \sigma^{-2} \sum_{i=1}^{n_j} \mathbf{f}(X_{ij})\, \mathbf{f}'(X_{ij}) = \sigma^{-2} F_j F_j';$$

$$Y_j = \sigma^{-2} \sum_{i=1}^{n_j} y_{ij} \mathbf{f}(X_{ij}) = \sigma^{-2} F_j \mathcal{Y}_j.$$

Очевидно, что $\mathbf{D}(\widehat{\Theta}_j) \leqslant \mathbf{D}(\widetilde{\Theta}_j)$, где $\widetilde{\Theta}_j$ — обычная мнк-оценка ($\widetilde{\Theta} = \mathbf{M}_j^{-1} \mathcal{Y}_j$). Выше предполагалось, что дисперсия σ^2 задана. В противном случае в (7.98) следует использовать какую-либо подходящую оценку этой величины, например

$$\widehat{\sigma^2} = \frac{1}{\sum_{j=1}^{k} (n_j - l)} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \Theta_j' \mathbf{f}(X_{ij}))^2.$$

Конечно, два последних требования из (7.97) будут при этом выполняться лишь приближенно. Данное замечание относится и к случаям, рассмотренным в п. 7.6.3 и 7.6.4. Нетрудно проверить, что оценки (7.98) могут быть получены так же, как решение следующей экстремальной задачи:

$$\widehat{\Theta}_j = \arg\min_{\Theta} \left[\sigma^2 \sum_{i=1}^{n_j} (y_{ij} - \Theta' \mathbf{f}(X_{ij}))^2 + (\Theta - \Theta_0)' \Sigma^{-1} (\Theta - \Theta_0) \right]. \tag{7.99}$$

7.6.3. Случай, когда известна только ковариационная матрица Σ (требуется оценить параметры Θ_j и Θ_0).

Начнем с оценки для Θ_0. Регрессионная задача (7.96) может быть переписана в виде

$$y_{ij} = (\Theta_0' \mathbf{f}(X_{ij}) + (\Theta_j - \Theta_0)' \mathbf{f}(X_{ij}) + \varepsilon_{ij} = \Theta_0' \mathbf{f}(X_{ij}) + \nu_{ij},$$

причем из свойств случайных величин Θ_j и ε_{ij} следует

$$\mathbf{E}\nu_{ij} = 0,\ \mathbf{E}(\nu_j \nu_{j'}') = \delta_{jj'} \Sigma_j;$$

$$\Sigma_j = \sigma^2 \mathbf{I}_{n_j} + F_j' \Sigma F_j,\ \nu_j' = (\nu_{1j}, \ldots, \nu_{n_j j}).$$

В соответствии с [15] (случай коррелированных наблюдений) наилучшие линейные оценки имеют следующий вид:

$$\widehat{\Theta}_0 = \mathcal{M}^{-1} \overline{Y},\ \mathbf{D}(\widehat{\Theta}_0) = \mathcal{M}, \tag{7.100}$$

где

$$\mathcal{M} = \sum_{j=1}^{k} \mathcal{M}_j;\ \overline{Y} = \sum_{j=1}^{k} \overline{Y}_j;$$

$$\mathcal{M}_j = \mathbf{F}_j \Sigma_j^{-1} \mathbf{F}_j';\ \overline{Y}_j = \mathbf{F}_j \Sigma_j^{-1} \mathcal{Y}_j.$$

Формула (7.100) приводит к весьма громоздким вычислениям, особенно при $n_j \gg l$, вследствие необходимости обращения матриц Σ_j, $j = \overline{1, k}$. Можно уменьшить объем вычислений, если прибегнуть к формуле

$$\Sigma_j^{-1} = \sigma^2 \mathbf{I}_{n_j} - \mathbf{F}_j' (\Sigma^{-1} + \sigma^{-2} \mathbf{F}_j \mathbf{F}_j')^{-1} \mathbf{F}_j,$$

которая является очевидным следствием известной формулы [117]

$$(\mathbf{A} + \mathbf{BCB}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{B}' \mathbf{A}^{-1} \mathbf{B} + \mathbf{C}^{-1})^{-1} \mathbf{B}' \mathbf{A}^{-1}.$$

При $|\mathbf{F}_j \mathbf{F}_j'| \neq 0$ удается добиться дальнейшего упрощения вычислений. Оказывается, что

$$\widehat{\Theta}_0 = \mathcal{M}^{-1} \sum_{j=1}^{k} (\mathbf{M}_j^{-1} + \Sigma)^{-1} \widetilde{\Theta}_j,$$

где $\widetilde{\Theta}_j = \mathbf{M}_j^{-1} Y_j$. Иными словами, $\widehat{\Theta}_0$ является линейной комбинацией наилучших линейных несмещенных оценок для каждой j-й серии наблюдений без учета случайного характера Θ_j. Это позволяет проводить основную часть расчетов по стандартным алгоритмам линейного регрессионного анализа.

Вычисления становятся совсем простыми, если $\mathbf{F}_j \equiv \mathbf{F}$, т. е. планы экспериментов над различными объектами одинаковы. При этом

$$\widehat{\Theta}_0 = k^{-1} \sum_{j=1}^{k} \widetilde{\Theta}_j, \quad \mathbf{D}(\widehat{\Theta}_0) = k^{-1} (\mathbf{M}^{-1} + \Sigma),$$

где $\mathbf{M} = \sigma^2 \mathbf{F}\mathbf{F}'$

Оценки $\widehat{\Theta}_j$ вычисляются по формуле (7.98) с заменой Θ_0 на оценку $\widehat{\Theta}_0$.

7.6.4. Случай неизвестных Θ_0 и Σ (требуется оценить Θ_j, Θ_0 и Σ). В качестве оценок $\widehat{\Theta}_j$ можно использовать (7.98), если Θ_0 и Σ заменить на любые подходящие оценки. Например, в качестве таких оценок можно выбрать следующие величины:

$$\widehat{\Theta}_0 = \lim_{s \to \infty} \Theta_{0s}, \quad \widehat{\Sigma} = \lim_{s \to \infty} \Sigma_s;$$

$$\Theta_{0s} = \mathcal{M}_s^{-1} \sum_{j=1}^{k} \mathcal{M}_{js} \widetilde{\Theta}_j, \quad \mathcal{M}_{js} = (\Sigma_s + \mathbf{M}_j^{-1})^{-1}; \tag{7.101}$$

$$\Sigma_s = (k-1)^{-1} \sum_{j=1}^{k} (\widetilde{\Theta}_j - \Theta_{0, s-1})) (\widetilde{\Theta}_j - \Theta_{0, s-1}) .$$

Если $\mathbf{F}_j \equiv \mathbf{F}$, то итерационная процедура (7.101) оказывается состоящей из одного шага:

$$\widehat{\Theta}_0 = k^{-1} \sum_{j=1}^{k} \widetilde{\Theta}_j, \quad \widehat{\Sigma} = (k-1)^{-1} \sum_{j=1}^{k} (\widetilde{\Theta}_j - \widehat{\Theta}_0)(\widetilde{\Theta}_j - \widehat{\Theta}_0)'.$$

ВЫВОДЫ

1. Общая математическая модель линейной регрессии имеет вид $Y = \mathbf{X}\Theta + \varepsilon$, где Y — $(n \times 1)$-вектор наблюдений, $\mathbf{X} = (X_1 \ldots X_n)'$ — $(n \times p)$-матрица плана экспериментов, X_k — регрессор k-го наблюдения, Θ — $(p \times 1)$ -вектор неизвестных параметров, ε — $(n \times 1)$ — вектор случайных ошибок. В классической постановке задачи линейной регрессии предполагается, что $\varepsilon \in N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, где $\mathbf{I}_n$ — $(n \times n)$-единичная матрица. Оценки по методу наименьших квадратов (мнк-оценки) отыскиваются из условия минимизации по Θ величины $||Y - \mathbf{X}\Theta||$. Когда $|\mathbf{X}'\mathbf{X}| \neq 0$ (ранг $\mathbf{X}$ равен p), $\widehat{\Theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$. Оценкой σ^2 является $s^2 = ||Y - \mathbf{X}\widehat{\Theta}|| / (n-r)$ где r — ранг матрицы $\mathbf{X}$. Случай, когда $\varepsilon \in N(\mathbf{0}, \sigma^2\mathbf{V})$, где $\mathbf{V}$— известная положительно определенная матрица, легко сводится к рассмотренному путем линейного преобразования Y и $\mathbf{X}$.

2. В классических предположениях в случаях, когда матрицу плана экспериментов можно представить состоящей из k взаимоортогональных совокупностей столбцов $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_k)$, $\mathbf{X}_i'\mathbf{X}_j = \mathbf{0}$, $i \neq j$, вычисления значительно упрощаются, и компоненты вектора $\Theta = (\Theta^{(1)'}, \ldots, \Theta^{(k)'})$, соответствующие $\mathbf{X}_h$, оцениваются независимо друг от друга. Для проверки гипотез $H_i: \Theta^{(i)} = \mathbf{0}$ (ранг $\mathbf{X}_i$ равен r_i) используются отношения $F = \widehat{\Theta}^{(i)'}\mathbf{X}_i'\mathbf{X}_i\widehat{\Theta}^{(i)}/r_i s^2$, имеющие, когда H_i верно, $F(r_i, n-r)$-распределение.

3. В классических предположениях мнк-оценки совпадают с оценками максимального правдоподобия и являются наилучшими среди всех несмещенных оценок Θ. Однако при отклонении распределения ε от нормального в сторону увеличения вероятности больших отклонений мнк-оценки быстро теряют свои оптимальные свойства. В связи с этим в практической работе широко используются функции потерь $\rho(u) \neq u^2$. Среди них выделяется функция $\rho_\lambda(u) = \lambda^{-1}(1 - \exp\{-\lambda u^2/2\})$, при $\lambda \to 0$ стремящаяся к $u^2/2$, а при $u \to \infty$ $(\lambda > 0)$ имеющая горизонтальную асимптоту. Она приводит к так называемым эв-оценкам параметров регрессионной зависимости (эв-регрессия или λ-регрессия). Эти оценки устойчивы к нарушению предположения нормаль-

ности, имеют наглядную геометрическую интерпретацию, для них (при весьма общих предположениях) получены асимптотические (при $n \to \infty$) разложения.

4. Формой учета априорных сведений о распределении параметров регрессионной модели является байесовское оценивание. При этом следует различать три подхода: частотный; стандартные рекомендации, как поступать в условиях неопределенности; субъективный. Частотный подход не вызывает возражений с методологических позиций. Во втором подходе априорная (несобственная) плотность распределения параметров полагается пропорциональной $d\theta^{(1)} \ldots d\theta^{(p)} d\sigma/\sigma$, что приводит порою к серьезным интерпретационным трудностям. Основная трудность субъективного подхода состоит в том, что информация, полученная из данных, рассматривается на равных основаниях с распределением, получаемым из не полностью формализованных соображений. Вместе с тем байесовское оценивание обладает замечательным свойством — если выборка разбита на две части, то эквивалентны результаты двух подходов к оцениванию:

1) применение байесовского оценивания к первой выборке, использование полученного апостериорного распределения в качестве априорного для второй и повторное байесовское оценивание параметров второй выборки;

2) одномоментное применение байесовского оценивания к объединенной выборке.

5. В экономических и технологических исследованиях при фиксированном значении регрессора X часто рассматривается многомерный отклик $Y = \mathbf{\Psi}'(X)\Theta + \boldsymbol{\varepsilon}$, где Y — $(l \times 1)$-вектор наблюдений при значении регрессора X, $\mathbf{\Psi}$ — известная $(l \times p)$-матричная функция X, Θ — $(p \times 1)$-вектор неизвестных параметров, а $\boldsymbol{\varepsilon}$ — $(l \times 1)$-вектор ошибок $\in N(0, \mathbf{V})$, где $\mathbf{V}$ — неизвестная положительно определенная $(l \times l)$-матрица. Оценка вектора в многомерной регрессии проводится одновременно с оценкой матрицы $\mathbf{V}$ путем итеративного решения нелинейной системы уравнений. Разработаны устойчивые методы оценки многомерной регрессии. Многомерная регрессия может использоваться при описании многомерных распределений.

6. Во многих задачах регрессионного типа разбиение переменных на две жесткие группы (в первую входят переменные, наблюдаемые с ошибкой, во вторую — переменные, значения которых известны точно) оказывается неадекватным реальному положению дел: все переменные наблюдаются или фиксируются с некоторыми ошибками. К настоящему времени в литературе предложен ряд моделей, описывающих подобные ситуа-

ции. Соответствующие им оценки базируются в основном на традиционном мнк.

7. При анализе поведения схожих объектов (например, реакция однородной группы больных на испытываемое лекарство) удобно использовать регрессионные модели второго рода (например, $y_{ij}=\Theta_j' \cdot \mathbf{f}(X_{ij})+\varepsilon_{ij}$, где индекс j соответствует номеру объекта). Предполагая, что параметры Θ_j (точнее, их изменчивость) могут быть описаны некоторой вероятностной моделью, удается построить оценки, которые оказываются эффективнее оценок, строящихся в отдельности для каждого j-го объекта без учета имеющейся информации о других схожих объектах.

В формальном плане эти оценки оказываются во многом схожи с байесовскими оценками.

Глава 8. ОЦЕНИВАНИЕ ПАРАМЕТРОВ РЕГРЕССИИ В УСЛОВИЯХ МУЛЬТИКОЛЛИНЕАРНОСТИ И ОТБОР СУЩЕСТВЕННЫХ ПРЕДИКТОРОВ

8.1. Явление мультиколлинеарности и его влияние на мнк-оценки

Рассмотрим обычную модель линейной по параметрам регрессии с неслучайными переменными $X=(x^{(1)}, \ldots, x^{(p)})'$:

$$y_i=\theta_0+X_i'\Theta+\varepsilon_i, \quad i=\overline{1, n}. \tag{8.1}$$

Оценки коэффициентов регрессии Θ получаются из решения системы уравнений (см. п. 8.6.1)

$$\mathbf{S}\widehat{\Theta}=\widehat{C}_{yX}, \tag{8.1'}$$

где $\mathbf{S}$ — матрица ковариаций объясняющих переменных размера $p \times p$, $\widehat{C}_{yX}$ — p-мерный вектор оценок ковариаций между объясняющими переменными и y.

Пусть теперь $\overline{X}=(\overline{x}^{(1)}, \ldots, \overline{x}^{(p)})'$ — вектор, компоненты которого суть средние значения предсказывающих переменных

$$\overline{x}^{(j)}=\frac{1}{n}\sum_{i=1}^{n} x_i^{(j)}. \tag{8.2}$$

Тогда с учетом очевидного тождества для свободного члена — $\theta_0=\mathbf{E}y-\overline{X}'\Theta$, его оценка может быть записана в виде $\widehat{\theta}_0=$

$= \bar{y} - \bar{X}' \hat{\Theta}$, где $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ — оценка среднего значения $\mathbf{E}y$.

Предсказанное значение $\hat{y}$ может быть вычислено по одной из следующих формул:

$$\hat{y}_i = \hat{\theta}_0 + X_i' \hat{\Theta} \text{ или } \hat{y}_i = \bar{y} + X_{ic}' \hat{\Theta}, \tag{8.3}$$

где $X_{ic} = X_i - \bar{X}$ — центрированный вектор X.

Матрица ковариаций между оценками параметров запишется

$$\mathbf{V}(\hat{\Theta}) = \frac{1}{n} \sigma^2 \mathbf{S}^{-1}, \tag{8.4}$$

а ее оценка

$$\hat{\mathbf{V}}(\hat{\Theta}) = \frac{1}{n} s^2 \mathbf{S}^{-1}, \tag{8.4'}$$

где s^2 — несмещенная оценка σ^2 (см. § 11.1).

Далее иногда будут использоваться и *стандартизованные* (нормированные) объясняющие переменные

$$(x^{(i)} - \bar{x}^{(i)})/\sigma_i, \tag{8.5}$$

где

$$\hat{\sigma}_i^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j^{(i)} - \bar{x}^{(i)})^2, \tag{8.5'}$$

$\hat{\sigma}_i^2$ — дисперсия переменной $x^{(i)}$.

Оценки коэффициентов регрессии для стандартизованных переменных получаются из решения системы уравнений

$$\mathbf{R}\hat{\Theta} = \hat{\sigma}_y \hat{\mathbf{r}}_{yX}, \tag{8.6}$$

где $\mathbf{R}$ — матрица корреляций объясняющих переменных, $\hat{\mathbf{r}}_{yX}$ — вектор оценок корреляций переменных X с y, $\hat{\sigma}_y^2 = \frac{1}{n} \sum_{j=1}^{n} (y_j - \bar{y})^2$.

Явление мультиколлинеарности возникает, если между объясняющими переменными существуют почти точные линейные зависимости (в интервале их изменения, определяемого матрицей плана $\mathbf{X}$). В случае существования точных линейных соотношений между переменными матрица $\mathbf{S}$ (а следовательно, и $\mathbf{R}$) будет вырожденной и значит обычная обратная матрица $\mathbf{S}^{-1}$ ($\mathbf{R}^{-1}$) не существует, а матрица $\mathbf{X}$ (мы рассматриваем

случай $n > p$) будет матрицей *неполного ранга*. (Случай точной линейной зависимости иногда называют «мультиколлинеарностью в строгом смысле»). В случае *почти* точных зависимостей матрицы **S** и **R** будут плохо обусловлены (см. п. 8.6).

Мультиколлинеарность в основном появляется в задачах пассивного эксперимента, когда исследователь, собирая данные, не может влиять на значения объясняющих переменных. В активном эксперименте матрица данных **X** планируется (см. [136]), причем таким образом, что либо матрица **S** хорошо обусловлена, либо априори точно известны линейные зависимости, имеющие место между строками (столбцами матрицы **X**), и, следовательно, ее ранг.

Применение обычного мнк в условиях мультиколлинеарности приводит к некоторым нежелательным последствиям (ниже используются нормированные переменные):

1) значения нормы вектора оценок параметров $\widehat{\Theta}$ и соответственно абсолютных величин отдельных его компонент могут быть очень велики; количественно оценить этот эффект можно, рассматривая величину среднего значения квадрата нормы вектора

$$\mathbf{E}\,\|\widehat{\Theta}\|^2 = \|\Theta\|^2 + \frac{\sigma^2}{n}\,\mathrm{Sp}\,\mathbf{R}^{-1} = \|\Theta\|^2 + \frac{\sigma^2}{n}\sum_{i=1}^{p}\frac{1}{\lambda_i}, \tag{8.7}$$

где λ_i $(i=\overline{1-p})$ собственные числа матрицы **R**; если минимальное собственное число $\lambda_{\min}$ достаточно мало, то вклад второго слагаемого будет велик;

2) дисперсии компонент вектора $\widehat{\Theta}$ могут стать относительно столь большими, что оценки параметров будут статистически незначимыми; из (11.11) легко получить, что дисперсия оценки параметра θ_i равна:

$$\mathbf{D}\widehat{\theta}_i = \frac{\sigma^2}{n\,(1-R_i^2)}, \tag{8.8}$$

где R_i^2 — коэффициент множественной корреляции между переменной $x^{(i)}$ и остальными предсказывающими переменными; сама оценка параметра $\widehat{\theta_i}$ распределена по нормальному закону $N\,(\theta_i,\ \mathbf{D}\widehat{\theta}_i)$ (см. (11.13)); очевидно, если $\theta_i/\sqrt{\mathbf{D}\hat{\theta}_i} \ll 1$, что может произойти при величине R_i^2, достаточно близкой к 1, то вероятность того, что значение $|\widehat{\theta}_i|$ превзойдет некоторый уровень, выбранный для отвержения нулевой гипотезы (т.е. гипотезы $\theta_i = 0$), будет мала;

3) абсолютные значения коэффициентов корреляции между оценками параметров $\widehat{\theta}_i$ и $\widehat{\theta}_j$ ($i, j = \overline{1, p}$; $i \neq j$) близки к 1, что делает, например, бессмысленным построение доверительных интервалов отдельно для каждой из этих оценок (в подобных ситуациях приходится строить *совместную* доверительную область для обеих оценок);

4) величины оценок $\widehat{\theta}_i$ существенно меняются при незначительном возмущении матрицы **X** (может измениться даже знак коэффициента $\widehat{\theta}_i$); здесь количественной характеристикой являются числа обусловленности матриц **X** и **R**

$$\varkappa(\mathbf{R}) = \lambda_{max}/\lambda_{min}; \quad \varkappa(\mathbf{X}) = \sqrt{\varkappa(\mathbf{R})}$$

(подробнее о числах обусловленности см, п.8.6).

Все эти эффекты затрудняют и без того сложную задачу интерпретации коэффициентов регрессии или вообще делают невозможным ее решение без привлечения новых способов обработки и дополнительной информации. В этих условиях нельзя применять уравнение регрессии и для прогноза значений переменной y. *В то же время если уравнение регрессии предполагается использовать для целей прогноза значений переменной y только в точках, близких к значениям объясняющих переменных $x^{(1)}, \ldots, x^{(p)}$ из матрицы данных* **X**, *то оно может оказаться вполне удовлетворительным: независимо от степени связи между предсказывающими переменными качество уравнения регрессии определяется значением коэффициента множественной корреляции $R_{y.X}$ между переменной y и переменными X (хотя при этом может быть необходимо принять некоторые предосторожности чисто вычислительного характера)*. Таким образом, последствия мультиколлинеарности тем серьезнее, чем больше информации мы хотим получить из имеющейся совокупности наблюдений.

8.2. Регрессия на главные компоненты

Поскольку мультиколлинеарность связана с высокой степенью корреляции между исходными переменными, можно попытаться обойти эту трудность, используя в качестве новых переменных некоторые линейные комбинации исходных переменных, выбранные так, чтобы корреляции между ними были малы или вообще отсутствовали. Тогда матрица корреляций между оценками параметров относительно новых переменных будет близка к диагональной, что существенно упростит интерпретацию.

Когда переменных немного или имеются некоторые априорные теоретические данные, выбор таких комбинаций может быть осуществлен из содержательных соображений; в более общей ситуации один из возможных подходов основывается на использовании так называемых главных компонент (см. [14, п. 10.5.2]), что приводит к *регрессии на главные компоненты* [195, 201, 219].

Пусть $U_1, \ldots, U_p$ — нормированные собственные векторы матрицы $\mathbf{R}$, расположенные в порядке убывания соответствующих им собственных чисел $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$. Тогда j-я главная компонента [14, п. 10.5.2] определяется как линейная комбинация исходных переменных, коэффициенты которой равны компонентам j-го собственного вектора, т. е. $z^{(j)} = \sum_{i=1}^{p} u_{ij} x^{(i)}$.

Поскольку главные компоненты некоррелированы, значения оценок $\widehat{g}_j$ параметров g_j регрессии при j-й компоненте не зависят от того, какие еще компоненты включены в уравнение регрессии, и равны:

$$\widehat{g}_j = \frac{1}{(n-1)\lambda_j} \sum_{k=1}^{n} z_k^{(j)} y_k = \frac{1}{\lambda_j} (\widehat{\mathbf{r}'_{yX}} U_j), \tag{8.9}$$

где $z_k^{(j)}$ — значение j-й главной компоненты для k-го наблюдения. Матрица ковариаций оценок $\widehat{g}_j$ диагональна, и непосредственно из (11.11) следует, что дисперсия j-го коэффициента $\widehat{g}_j$ равна:

$$\mathbf{D}\widehat{g}_j = \sigma^2/(n-1)\lambda_j, \tag{8.10}$$

т. е. ошибка коэффициента регрессии минимальна для первой главной компоненты и растет с увеличением номера главной компоненты.

Квадрат коэффициента корреляции между j-й главной компонентой и y

$$\widehat{r}_j^2 = (\widehat{\mathbf{r}'_{yX}} U_j)^2 / \lambda_j \widehat{\sigma}_y^2. \tag{8.11}$$

Отсутствие корреляции между главными компонентами позволяет легко организовать пошаговую процедуру отбора (см. п.8.7.3) информативных для предсказания y главных компонент, результат которой в этом случае будет эквивалентен полному перебору.

Рассмотрим следующие критерии отбора, использующие главные компоненты.

1. **t-статистика для проверки значимости коэффициента регрессии при j-й главной компоненте:**

$$t_j = \frac{(n-1)\lambda_j}{s} \hat{g}_j. \tag{8.12}$$

В случае истинности нулевой гипотезы ($g_j = 0$) эта величина имеет t-распределение. Будем использовать схему пошагового удаления переменных. Задаваясь некоторым пороговым значением $t_{\text{удал}}$, исключаем из уравнения регрессии j-ю главную компоненту, если

$$|t_j| < t_{\text{удал}}. \tag{8.13}$$

В силу независимости оценок параметров $\hat{g}_j$ никакого пересчета остальных коэффициентов при удалении той или иной главной компоненты проводить не надо. Обычно в качестве $t_{\text{удал}}$, выбирают значения $t_{0,1}$, $t_{0,05}$, $t_{0,025}$ для t-распределения с соответствующим числом степеней свободы. Другой способ выбора критического значения дан в п. 8.5.2 (см. (8.58)).

Число степеней свободы зависит от того, какая оценка дисперсии ошибки используется. Можно использовать мнк-оценку дисперсии или, что эквивалентно, оценку дисперсии, получаемую при включении в уравнение регрессии всех p главных компонент. Тогда число степеней свободы $\nu = n - p - 1$, а оценка дисперсии s^2 имеет вид:

$$s^2 = \frac{\hat{\sigma}_y^2 (1 - \hat{R}_{y.p}^2)}{n-p-1}, \tag{8.14}$$

где $\hat{R}_{y \cdot p} = \hat{R}_{y \cdot X}$ — оценка коэффициента множественной корреляции между y и всеми p главными компонентами.

С другой стороны, пусть после удаления очередной главной компоненты j_{k+1} осталось k главных компонент. Тогда, продолжая процедуру отбора, можно использовать оценку дисперсии s_k^2, соответствующую уравнению с *оставшимися k главными компонентами*:

$$s_k^2 = \frac{\hat{\sigma}_y^2 (1 - \hat{R}_{y.k}^2)}{n-k-1}, \tag{8.15}$$

где $\hat{R}_{y \cdot k}$ — коэффициент множественной корреляции между y и оставшимися k главными компонентами. Поскольку главные компоненты некоррелированы, то имеем $R_{y \cdot k}^2 = \sum_{i=1}^{p} \delta_i r_i^2$, где $\delta_i = 1$, если главная компонента включена в набор главных

компонент, входящих в уравнение регрессии, и 0 — в противном случае. При такой оценке дисперсии число степеней свободы $\nu = n - k - 1$.

2. **F-статистика для добавочной информации.** Используем пошаговую процедуру простого присоединения главных компонент. Пусть в наборе уже имеется k главных компонент. Тогда из всех оставшихся главных компонент находим компоненту с максимальным значением F-статистики

$$F = \frac{R^2_{y.\,k+1} - R^2_{y.k}}{1 - R^2_{y\ k+1}}(n-k-2)$$

и включаем ее в уравнение регрессии, если выполняется условие $F > F_{\text{вкл}}$. В качестве критического значения $F_{\text{вкл}}$ берут значения процентных точек, например $F_{0{,}05}$, $F_{0{,}025}$ для F-распределения с одной и $\nu = n - k - 2$ степенями свободы. Если компонент, для которых выполняется условие $F > F_{\text{вкл}}$, нет, то процесс отбора главных компонент считается оконченным. Можно показать, что использование F-критерия приводит к тому же набору компонент, что и использование t-критерия с меняющейся оценкой дисперсии (8.15).

3. **Величина собственного числа для i-й главной компоненты.** Именно эта величина предлагается для отбора главных компонент в некоторых работах [163, 43, 219]. Если $x^{(1)}, \ldots, x^{(p)}$ сильно взаимно коррелируют, то, начиная с некоторого номера i_0, значения собственных чисел $\lambda_{i_0+1}, \ldots, \lambda_p$ близки к нулю, а соответствующие коэффициенты регрессии могут стать большими по абсолютной величине. Дисперсии оценок коэффициентов регрессии, соответствующих этим главным компонентам, также будут велики. Отсюда следует целесообразность удаления главных компонент с малыми собственными числами, т. е. полагаем

$$\tilde{g}_j = \begin{cases} \hat{g}_j, & \text{если } \lambda_j > \lambda_{\text{кр}}; \\ 0, & \text{если } \lambda_j \leqslant \lambda_{\text{кр}}, \end{cases} \qquad (8.16)$$

или, учитывая, что главные компоненты упорядочены по убыванию собственных чисел,

$$\tilde{g}_j = \begin{cases} \hat{g}_j, & \text{если } j < i_0; \\ 0, & \text{если } j \geqslant i_0, \end{cases} \qquad (8.16')$$

где i_0 — первый номер, для которого выполняется неравенство $\lambda_{i_0} < \lambda_{\text{кр}}$.

Критическое значение $\lambda_{кр}$ обычно выбирается в виде

$$\lambda_{кр} = e\,\mathrm{Sp}(\mathbf{R}) = e \sum_{j=1}^{p} \lambda_j, \qquad (8.17)$$

где $\mathrm{Sp}(\mathbf{R}) = p$ — след корреляционной матрицы; e — малая величина, например 10^{-5}.

Другой метод выбора числа компонент основан на общепринятой методологии использования главных компонент. Задаемся некоторой величиной доли следа α, близкой к 1, и включаем в уравнение регрессии компоненты до тех пор, пока

$$\sum_{j=1}^{i} \lambda_j / \mathrm{Sp}(\mathbf{R}) < \alpha. \qquad (8.18)$$

Как только это неравенство перестает выполняться, включение компонент прекращается, и коэффициенты регрессии оставшихся главных компонент объявляются статистически незначимыми.

Подход к отбору главных компонент на основе величины собственных чисел эквивалентен регуляризации при вычислении псевдообратной матрицы на ЭВМ [17]. Он может быть использован и при наличии точной линейной зависимости между переменными, которая, однако, «замаскирована» ошибками округления при представлении данных в ЭВМ.

Однако процедуры отбора главных компонент, основанные на t-и F-статистиках, правильнее нацелены на решение сущности задачи, хотя при их использовании *могут быть отброшены и некоторые главные компоненты, соответствующие большим значениям* λ_i (если они слабо коррелированы с переменной y). Правда, как правило, компоненты с малыми значениями собственных чисел оказываются одновременно и слабо коррелированными с y и также отбрасываются, так что отбор существенных главных компонент по этим критериям автоматически приводит и к регуляризации задачи. Зная включенные в уравнение компоненты и соответствующие им коэффициенты регрессии, легко найти коэффициенты регрессии относительно исходных переменных $x^{(1)}, \ldots, x^{(p)}$

$$\widehat{\theta}_i = \sum_{k=1}^{p} \delta_k \widehat{g}_k u_{ik}, \qquad (8.19)$$

где $\delta_k = 1$, если главная компонента включена в информативный набор, и $\delta_k = 0$ — в противном случае.

Вообще говоря, полученные таким образом оценки для коэффициентов θ_i будут *смещенными*. Формулы для дисперсий и смещений этих коэффициентов приведены в п. 8.5.

8.3. Смещенное оценивание коэффициентов регрессии

Как известно (см. п. 7.1.2 и 11.1.1), мнк-оценки являются несмещенными оценками с минимальной дисперсией в классе линейных по $Y = (y_1, \ldots, y_n)'$ оценок. Однако в условиях мультиколлинеарности эта минимальная дисперсия может быть чрезмерно велика. Оказывается, *если отказаться от несмещенности*, можно построить линейные по Y оценки $\widetilde{\Theta}$, для которых средний квадрат отклонения от истинных значений параметров Θ будет меньше, чем для мнк-оценок $\widehat{\Theta}$, т. е.

$$\mathbf{E}(\Theta-\widetilde{\Theta})'(\Theta-\widetilde{\Theta}) < \mathbf{E}(\Theta-\hat{\Theta})'(\Theta-\hat{\Theta}). \qquad (8.20)$$

Любую оценку $\widetilde{\Theta}$, линейную по Y, можно представить в виде

$$\widetilde{\Theta} = \mathbf{C}\widehat{\Theta}, \qquad (8.21)$$

где $\widehat{\Theta}$ обычная мнк-оценка, а $\mathbf{C}$ — матрица размера $p \times p$, не обязательно невырожденная, называемая матрицей редукции.

Оценка вида (8.21) имеет следующие математическое ожидание и матрицу ковариаций:

$$\mathbf{E}\widetilde{\Theta} = \mathbf{C}\Theta; \quad \mathbf{V}(\widetilde{\Theta}) = \mathbf{C}\cdot\mathbf{V}(\widehat{\Theta})\cdot\mathbf{C}' = \frac{\sigma^2}{n}\,\mathbf{C}\mathbf{S}^{-1}\mathbf{C}'. \qquad (8.22)$$

Для нормированной суммы квадратов отклонений имеем

$$\widehat{\Delta}_n(\widetilde{\Theta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y}-(\widetilde{\Theta}'X_c))^2 = \widehat{\Delta}(\widehat{\Theta}) + (\widetilde{\Theta}-\widehat{\Theta})'\,\mathbf{S}\,(\widetilde{\Theta}-\widehat{\Theta}), \qquad (8.23)$$

где $\widehat{\Delta}_n(\widehat{\Theta})$ — нормированная сумма квадратов отклонений для мнк-оценки. После некоторых преобразований выражение (8 23) можно записать:

$$\widehat{\Delta}_n(\widetilde{\Theta}) = \widehat{\Delta}_n(\widehat{\Theta}) + \widehat{\Theta}'(\mathbf{C}-\mathbf{I}_p)'\,\mathbf{S}\,(\mathbf{C}-\mathbf{I}_p)\,\widehat{\Theta}. \qquad (8.24)$$

Среднее значение величины $\widehat{\Delta}_n$ равно:

$$\mathbf{E}\widehat{\Delta}_n(\widetilde{\Theta}) = \frac{(n-p-1)}{n}\sigma^2 + \Theta'(\mathbf{C}-\mathbf{I}_p)S(\mathbf{C}-\mathbf{I}_p)\Theta + \frac{1}{n}\mathrm{Sp}(\mathbf{C}-\mathbf{I}_p)^2. \tag{8.25}$$

Введем функционал, характеризующий качество оценки (8.21) (функцию потерь)

$$L^2_{\mathbf{W}}(\widetilde{\Theta}, \Theta) = \mathbf{E}[(\Theta - \widetilde{\Theta})'\mathbf{W}(\Theta - \widetilde{\Theta})], \tag{8.26}$$

где $\mathbf{W}$ — неотрицательно определенная весовая матрица.

Наиболее часто используются весовые матрицы вида $\mathbf{W} = \mathbf{I}_p$, $\mathbf{W} = \mathrm{diag}(s_{11}, \ldots, s_{pp})$, $\mathbf{W} = \mathbf{S}$.

Будем искать теперь оценки Θ, минимизирующие функцию потерь (8.26).

$$\widetilde{\Theta} = \mathbf{C}^*\widehat{\Theta},\ \mathbf{C}^* = \arg\min_{\mathbf{C}} L^2_{\mathbf{W}}(\mathbf{C}\widehat{\Theta}, \Theta). \tag{8.27}$$

Такие оценки допускают следующую интерпретацию. Пусть, используя матрицу $\mathbf{X}$ и n-мерный вектор значений прогнозируемой величины Y, мы получили некоторую оценку параметров $\widetilde{\Theta}$ и среднего значения зависимой переменной $\bar{y}$. Используем теперь эти оценки для прогноза значений переменной y для векторов X^*, не входящих в матрицу $\mathbf{X}$. Будем считать при этом, что модель (8.1) остается верной, а компоненты векторов X^* распределены согласно некоторому закону распределения с вектором средних значений $\bar{X} = (\bar{x}^{(1)}, \ldots, \bar{x}^{(p)})'$ и матрицей ковариаций $\mathbf{W}$. Пусть $\hat{\delta}^2(X^*, \widetilde{\Theta})$ есть квадрат ошибки предсказания значения y^* для вектора X^*:

$$\widehat{\delta}^2(X^*, \widetilde{\Theta}) = (y^* - \widetilde{\Theta}'(X^* - \bar{X}) - \bar{y})^2 = (\Theta' X^*_c + \mathbf{E}y + \varepsilon^* - \widetilde{\Theta}' X^*_c - \bar{y})^2,$$

где $X^*_c = X^* - \bar{X}$ — центрированный вектор X^*.

Тогда

$$\mathbf{E}_{X^*}\delta^2(X^*, \widetilde{\Theta} \mid Y) = (\Theta - \widetilde{\Theta})'\mathbf{W}(\Theta - \widetilde{\Theta}) + (y - \mathbf{E}\bar{y})^2 + (\varepsilon^*)^2 + 2\varepsilon^*(\bar{y} - \mathbf{E}y).$$

Усредняя теперь по ε^*, имеем

$$\mathbf{E}_{\varepsilon^*}\mathbf{E}_{X^*}\widehat{\delta}^2(X^*, \widetilde{\Theta} \mid Y) = (\Theta - \widetilde{\Theta})'\mathbf{W}(\Theta - \widetilde{\Theta}) + \sigma^2 + (\bar{y} - \mathbf{E}y)^2.$$

Взяв далее математическое ожидание по Y, получаем, что

$$\mathbf{E}\widehat{\delta}^2(X^*, \widetilde{\Theta}) = L^2_{\mathbf{W}} + \frac{\sigma^2}{n} + \sigma^2. \tag{8.28}$$

Таким образом, *уравнение регрессии с параметрами, определенными из условия минимума функционала (8.26), минимизирует математическое ожидание квадрата ошибки прогноза на векторах X^*, не входящих в состав матрицы плана* **X**, *использованной для оценки, в то время как обычная мнк-оценка минимизирует сумму квадратов отклонений для матрицы* **X**.

Линейное преобразование объясняющих переменных. Рассмотрим теперь, как преобразуются оценки параметров уравнения регрессии и функционал (8.26) при линейном преобразовании объясняющих переменных.

Пусть для некоторого набора переменных $x^{(1)},\ldots,\ x^{(p)}$ определена оценка вида (8.21), удовлетворяющая условию (8.27) минимума функционала (8.26) с весовой матрицей $\mathbf{W}_X$. Перейдем теперь к системе переменных $z = (z^{(1)},\ldots,\ z^{(p)})'$, связанных с X невырожденным линейным преобразованием $\mathbf{L} : Z = \mathbf{L}'X$. Тогда мнк-оценкой параметров уравнения регрессии для переменных Z будет вектор

$$\widehat{\Theta}_Z = \mathbf{L}^{-1}\widehat{\Theta}_X, \qquad (8.29)$$

где через $\widehat{\Theta}_X$, $\widehat{\Theta}_Z$ обозначена мнк-оценка соответственно для переменных X (Z).

Аналогично смещенная оценка $\widetilde{\Theta}_X$ (8.21) преобразуется в оценку

$$\widetilde{\Theta}_Z = \mathbf{L}^{-1}\widetilde{\Theta}_X = \mathbf{L}^{-1}\mathbf{C}_X\widehat{\Theta}_X. \qquad (8.29)$$

С учетом (8.29) имеем

$$\widetilde{\Theta}_Z = (\mathbf{LCL}^{-1})\,\widehat{\Theta}_Z. \qquad (8.30)$$

Таким образом, оценке параметров уравнения регрессии в пространстве переменных X с матрицей редукции $\mathbf{C}_X$ в пространстве переменных Z соответствует оценка с матрицей редукции

$$\mathbf{C}_Z = \mathbf{L}\mathbf{C}_X\mathbf{L}^{-1}. \qquad (8.31)$$

Весовая матрица в мере качества оценки тоже меняется. Имеем

$$L^2_{\mathbf{W}_X}(\widetilde{\Theta}_X,\ \Theta_X) = \mathbf{E}\,(\widetilde{\Theta}_X - \Theta_X)'\,\mathbf{W}_X\,(\widetilde{\Theta}_X - \Theta_X) =$$
$$= \mathbf{E}\,(\widetilde{\Theta}_Z - \Theta_Z)\,(\mathbf{L}^{-1})'\,\mathbf{W}_X\,\mathbf{L}^{-1}\,(\widetilde{\Theta}_Z - \Theta_Z) = L^2_{\mathbf{W}_Z}(\widetilde{\Theta}_Z,\ \Theta_Z). \qquad (8.32)$$

Таким образом, матрица $\mathbf{C}_Z$ получается как решение задачи минимизации функционала (8.26) с *преобразованной* весовой матрицей

$$\mathbf{W}_X \to \mathbf{W}_Z = (\mathbf{L}^{-1})' \mathbf{W}_X \mathbf{L}^{-1}. \qquad (8.33)$$

Заметим, что если $\mathbf{W}_X$ есть ковариационная матрица пере менных X, то матрица $\mathbf{W}_Z$ будет ковариационной матрицей переменных Z.

8.4. Редуцированные оценки для стандартной модели линейной регрессии

Как уже указано в § 8.3, общий вид редуцированной оценки коэффициентов регрессии задается с помощью соотношения (8.21). Используемая там матрица редукции $\mathbf{C}$, как показано дальше, является либо функцией неизвестных параметров Θ, т. е. $\mathbf{C} = \mathbf{C}(\Theta)$, либо функцией оценок этих параметров $\mathbf{C} = \mathbf{C}(\widehat{\Theta})$. Следовательно, в последнем случае $\mathbf{C}$ будет случайной величиной. Такую матрицу назовем стохастической (формулы (8.22), (8.25) в случае стохастической матрицы уже не будут верными). Если матрица $\mathbf{C}$ — стохастическая, то оценки вида (8.21), строго говоря, не будут линейными по Y, однако самое важное их свойство, определяющее их полезность для приложений,— уменьшение среднего квадрата отклонений (8.26) (в метрике матрицы $\mathbf{W}$) — сохраняется.

Первоначально название «редуцированные («shrinkage») оценки» относилось к оценкам вида $\widetilde{\Theta} = \lambda\widehat{\Theta}$, где скаляр $\lambda \in (0,1)$. Матрица $\mathbf{C}$ для этой оценки имеет вид $\mathbf{C} = \mathrm{diag}(\lambda, \ldots, \lambda)$. Смысл введения множителя λ состоит в уменьшении длины (евклидовой нормы) вектора оценок $\widetilde{\Theta}$, по сравнению с $\widehat{\Theta}$, которая в условиях мультиколлинеарности может существенно превышать длину истинного вектора параметров Θ (см. (8.7)).

8.4.1. Оценка Джеймса — Стейна. Для рассмотрения оценки Джеймса — Стейна перейдем предварительно к ортонормированным переменным $V = (v^{(1)}, \ldots, v^{(p)})'$ и модель регрессии запишем в виде

$$y_i = \gamma_0 + \Gamma' V_i + \varepsilon_i, \; i = \overline{1, n}, \qquad (8.34)$$

и $\mathbf{V}\mathbf{V}' = \mathbf{I}_p$, $\overline{V} = 0$.

Такая модель может быть получена, например, в полиномиальной регрессии при переходе к ортонормированной системе полиномов. В общей модели регрессии ортонормированными

переменными, в частности, будут переменные $v^{(i)} = z^{(i)}/\sqrt{n\lambda_j}$, где $z^{(i)}$ — главные компоненты (см. § 8.2) матрицы $\mathbf{X}$.

Мнк-оценка для коэффициентов Г записывается в виде $\widehat{\Gamma} = n\widehat{C}_{yV}$, и ее распределение подчиняется p-мерному нормальному закону $\widehat{\Gamma} \sim N_p(\Gamma, \sigma^2\mathbf{I}_p)$.

Пусть теперь в качестве функции потерь, соответствующей некоторой оценке $\widetilde{\Gamma}$ параметров регрессии Г, используется функция потерь вида (8.26) с *единичной матрицей*, т.е.

$$L_{\mathrm{I}}^2(\widetilde{\Gamma}, \Gamma) = \sum_{i=1}^{p} \mathbf{E}(\widetilde{\Gamma}_i - \Gamma_i)^2 = \mathbf{E}\,\|\widetilde{\Gamma} - \Gamma\|^2. \qquad (8.35)$$

Для мнк-оценки $L_{\mathrm{I}}^2(\widehat{\Gamma}, \Gamma) = p\sigma^2$, верна следующая теорема [216].

Теорема Джеймса — Стейна. Пусть $p \geqslant 3$. Тогда оценка

$$\widetilde{\Gamma}(c) = \lambda_{JS}\widehat{\Gamma}, \qquad (8.36)$$

где $\lambda_{JS} = (1 - nc\widehat{\Delta}_n(\widehat{\Gamma})/\|\widehat{\Gamma}\|^2)$, c — любое число в интервале $0 < c < 2(p-2)/(n-p+2)$, «лучше» мнк-оценки $\widehat{\Gamma}$, в смысле критерия (8.35), каков бы ни был вектор неизвестных параметров Г. Иными словами, при *любом* Г верно неравенство

$$L_{\mathrm{I}}^2(\widetilde{\Gamma}(c), \Gamma) = \mathbf{E}\,\|\widetilde{\Gamma}(c) - \Gamma\|^2 \leqslant \mathbf{E}\,\|\widehat{\Gamma} - \Gamma\|^2 = p\sigma^2.$$

Условие $p \geqslant 3$ является существенным, так как, как показано в [216], когда $p = 1$ или $p = 2$, не существует оценки Γ^* лучшей, чем мнк-оценка в смысле (8.35), т. е. такой оценки, чтобы $L_{\mathrm{I}}^2(\Gamma^*, \Gamma) < L_{\mathrm{I}}^2(\widehat{\Gamma}, \Gamma)$ для всех Г.

Используя оценку коэффициента множественной корреляции между y и X, множитель Стейна можно записать в виде, инвариантном относительно преобразования предсказывающих переменных

$$\lambda_{JS} = 1 - c\,\frac{n-p-1}{n}\,(1 - \widehat{R}^2_{y.X})/\widehat{R}^2_{y.X}. \qquad (8.37)$$

Когда $c = 2(p-2)/(n-p+1)$, получим оценку $\widetilde{\Gamma}$, для которой $L_{\mathrm{I}}^2(\widetilde{\Gamma}, \Gamma) = L_{\mathrm{I}}^2(\widehat{\Gamma}, \Gamma) = p\sigma^2$ при всех Г, так что это значение c приводит к оценке, не лучшей чем мнк-оценка. Если $c = 0$, оценка Стейна, очевидно, просто совпадает с мнк-оценкой. Минимальное значение функции потерь $L_{\mathrm{I}}^2(\widetilde{\Gamma}(c), \Gamma)$

достигается при значении $c^* = (p-2)/(n-p+2)$. Тогда $L_1^2(\tilde{\Gamma}(c^*), \Gamma) = 2n\sigma^2/(n-p+2)$, т. е. примерно равно $2\sigma^2$, когда $n \gg p$. Отсюда следует, что оценка Джеймса — Стейна при больших p и n лучше мнк-оценки примерно в $p/2$ раз.

В то же время при наличии мультиколлинеарности оценка Джеймса — Стейна может оказаться *столь же неудовлетворительной, как и обычная мнк-оценка*. Чтобы показать это, вернемся от ортонормированных переменных V к главным компонентам Z, что соответствует линейному преобразованию $z^{(j)} = \sqrt{n\lambda_j}v^{(j)}$ $(j = \overline{1, p})$. Тогда согласно формуле (8.30) оценка Джеймса — Стейна для параметров уравнения регрессии на главные компоненты будет иметь в точности вид (8.36). т. е.

$$\tilde{G} = \lambda_{JS}\hat{G}. \tag{8.38}$$

Однако согласно формуле (8.32) оценка $\tilde{G}$ минимизирует уже не функцию потерь (8.35), а функцию потерь

$$L_\lambda^2(\tilde{G}, G) = \mathbf{E}\sum_{i=1}^{p}\lambda_i(\tilde{g}_i - g_i)^2. \tag{8.39}$$

Таким образом, ошибки оценок коэффициентов $\tilde{g}_i$, соответствующих главным компонентам с минимальными значениями дисперсии λ_i, т. е. компонентам, «наиболее ответственным» за мультиколлинеарность, входят в функцию потерь с минимальными весами λ_i. Это означает, что улучшение оценки Джеймса — Стейна по сравнению с мнк-оценкой достигается в первую очередь за счет уменьшения вклада компонент с относительно большой дисперсией, хотя при мультиколлинеарности, напротив, следует подавлять вклад компонент с минимальной дисперсией.

Улучшенная оценка Джеймса — Стейна. Как следует из выражения (8.37), при достаточно малых значениях $R_{y\,X}$ множитель λ_{JS} может стать отрицательным. Этого недостатка лишена улучшенная оценка типа Джеймса — Стейна, приведенная в [249] [1]. Она определяется как редуцированная оценка

$$\tilde{\Theta}^* = \mu\hat{\Theta}. \tag{8.40}$$

[1] В [249] приведены ссылки на источники, в которых получена и изучалась улучшенная оценка Джеймса—Стейна.

где множитель

$$\mu(c) = \begin{cases} \lambda_{JS}(c), & \text{когда } \lambda_{JS}(c) > 0; \\ 0, & \text{когда } \lambda_{JS}(c) \leqslant 0, \end{cases} \tag{8.41}$$

$\widehat{\Theta}$ — обычная мнк-оценка.

Для ортонормированных переменных V показано [249], что оценка с редуцирующим множителем μ лучше оценки Джеймса — Стейна (а тем более мнк-оценки по критерию L_I^2 $(\widetilde{\Gamma}, \Gamma)$), хотя оптимальное значение c^* и соответствующее минимальное значение L_I^2 $(\widetilde{\Gamma}(c^*), \Gamma)$ для нее аналитически не определены. Однако можно полагать, что они близки соответствующим значениям для оценки Джеймса — Стейна.

Для регрессии y на главные компоненты и на исходные переменные оценки типа (8.40) лучше оценки Джеймса — Стейна и мнк-оценки по соответственно взвешенным критериям L_Λ^2 и L_S^2.

Применение оценки Джеймса — Стейна для уточнения части параметров. Оценку Джеймса — Стейна, равно как и улучшенную оценку (8.40), можно применить для уточнения части параметров уравнения регрессии, лишь бы количество уточняемых параметров q удовлетворяло неравенству $q \geqslant 3$.

Рассмотрим снова модель (8.34). Представим вектор Γ' в виде $\Gamma' = (\Gamma^{(1)'}, \Gamma^{(2)'})$, где $\Gamma^{(1)'}$ имеет размерность $p-q$, а $\Gamma^{(2)'}$ — размерность q. Вектор $\widehat{\Gamma}$ разобьется на два подвектора $\widehat{\Gamma}' = (\widehat{\Gamma}^{(1)'}, \Gamma^{(2)'})$ размерности $p-q$ и q соответственно Введем множитель

$$\lambda_{JS}^{(2)} = 1 - c^{(2)} \frac{\widehat{\Delta}_n(\widehat{\Gamma})}{\|\widehat{\Gamma}^{(2)}\|^2},$$

где $0 < c^{(2)} < 2(q-2)/(n-p+2)$.
Тогда оценки

$$\widetilde{\Gamma} = \begin{pmatrix} \widehat{\Gamma}^{(1)} \\ \lambda_{JS}^{(2)} \widehat{\Gamma}^{(2)} \end{pmatrix}, \quad \widetilde{\widetilde{\Gamma}} = \begin{pmatrix} \widehat{\Gamma}^{(1)} \\ \mu^{(2)} \widehat{\Gamma}^{(2)} \end{pmatrix} \tag{8.41), (8.42}$$

лучше мнк-оценки по критерию (8.35). Оптимальное значение $c^{(2)} = (q-2)/(n-p+2)$. В таком виде оценка Джеймса— — Стейна позволяет существенно улучшить мнк-оценку в условиях мультиколлинеарности. Действительно, выделяя во вторую составляющую $\widehat{\Gamma}^{(2)}$ вектора $\widehat{\Gamma}$ коэффициенты, соответствующие, например, малым собственным числам λ_i или ма-

лым значениям t-статистики (8.12), и используя затем множитель Стейна, можно существенно уменьшить вклад этих компонент в оценку параметров уравнения регрессии при возвращении к исходным переменным.

8.4.2. Редуцированная оценка Мейера — Уилке. Матрица редукции **C** для этой оценки получается как решение задачи минимизации следа ковариационной матрицы вектора $\widetilde{\Theta} = \mathbf{C}^*\Theta$ при условии, что нормированная сумма квадратов отклонений $\widehat{\Delta}_n(\widetilde{\Theta}) = \delta_0 > 0$ [231]. Используя формулы (8.22), (8.24), задачу минимизации для определения матрицы $\mathbf{C}^*$ можно записать в виде

$$\mathbf{C}^* = \arg\min_{\mathbf{C}} \mathrm{Sp}\,(\mathbf{C}\mathbf{S}^{-1}\mathbf{C}') \qquad (8.43)$$

$$\text{при условии } \widehat{\Delta}_n(\widehat{\Theta}) + \widehat{\Theta}'(\mathbf{C}^* - \mathbf{I})\,\mathbf{S}\,(\mathbf{C}^* - \mathbf{I})\,\widehat{\Theta} = \delta_0, \qquad (8.43')$$

что дает в результате

$$\mathbf{C}^* = \delta\widehat{\Theta}\widehat{\Theta}'(\mathbf{I} + \delta\widehat{\Theta}\widehat{\Theta}')^{-1}, \qquad (8.44)$$

где δ выбирается так, чтобы выполнялось условие (8.43′), откуда после преобразования по формуле Бартлетта [117] оценка[1] запишется

$$\widetilde{\Theta} = \lambda_{MW}\,\widehat{\Theta}, \qquad (8.45)$$

где

$$\lambda_{MW} = \frac{\delta\,\|\widehat{\Theta}\|^2}{1 + \delta\,\|\widehat{\Theta}\|^2}. \qquad (8.46)$$

Как положительное качество оценки (8.45) отметим, что множитель λ_{MW} является функцией только мнк-оценки. С другой стороны, поскольку оценка Мейера и Уилке является стохастической редуцированной оценкой, формула (8.22) для ковариационной матрицы будет неверна (матрица $\mathbf{C}^*\mathbf{S}^{-1}\mathbf{C}^{*\prime}$ отнюдь не является в этом случае ковариационной матрицей оценки), поэтому нельзя утверждать, как это делают авторы оценки, что она минимизирует след ковариационной матрицы. Величина функционала качества (8.26) для нее также пока неизвестна, так что в отличие от оценки Стейна нельзя сказать, при каких условиях и в каком смысле она лучше мнк-оценки.

Некоторые другие типы редуцированных оценок приведены в [43, § 6.5].

---

[1] В [231] при преобразованиях была допущена ошибка, в результате чего множитель λ_{MW} определен неверно. Эта ошибка исправлена в [43].

8.5. Оценки, связанные с ортогональным разложением

Использование функционала L_W^2 (8.26) как меры качества оценки $\tilde{\Theta}$ *не гарантирует* еще, что каждая компонента вектора $\tilde{\Theta}$ имеет меньшую среднеквадратическую ошибку, чем вектор мнк-оценок (см., например, п. 8.3.1). Однако, как показано в [192], оценки, уменьшающие среднеквадратическую ошибку каждой из компонент вектора, существуют, в частности такими являются ридж-оценки [208, 209]. В настоящем разделе проводится рассмотрение достаточно общего класса оценок, обладающих указанным выше свойством.

Вернемся к регрессии на главные компоненты $Z = (z^{(1)},\ldots, z^{(p)})$ (см. п. 8.2). Пусть, как и прежде, вектор G есть вектор теоретических значений параметров, а $\hat{G}$ — вектор мнк-оценок. Пусть $U_1,\ldots, U_p$ — собственные векторы матрицы $\mathbf{S}$. Поскольку матрица $\mathbf{S}$ невырождена, векторы U_i $(i = \overline{1, p})$ образуют полную ортонормированную систему (см. [102]), и поэтому любой вектор оценок параметров может быть представлен в виде

$$\tilde{\Theta} = \sum_{i=1}^{p} \tilde{g}_i U_i. \qquad (8.47)$$

Для мнк-оценки $\tilde{g}_i = \hat{g}_i = \frac{\hat{C}'_{yX} U_i}{\lambda_i}$, где $\hat{C}_{yX}$ — оценка вектора ковариаций C_{yX} между прогнозируемой переменной y и переменными $(x^{(1)},\ldots, x^{(p)})$. Для самого вектора неизвестных параметров $\tilde{g}_i = g_i$. Мы будем рассматривать класс оценок вида (8.47) с коэффициентом $\tilde{g}_i = a_i\hat{g}_i$.

Таким образом, множитель a_i можно рассматривать как *относительный вес i-й главной компоненты в оценке параметров регрессии* $\tilde{\Theta}$ (8.47) *по сравнению с ее весом в мнк-оценке.*

Дальше веса $A = (a_1,\ldots, a_p)$ будут определяться из условия минимума функционала качества (8.26). Будем полагать при этом, что весовая матрица $\mathbf{W}$ перестановочна с матрицей $\mathbf{S}$, т. е. что векторы U_i являются и собственными векторами матрицы $\mathbf{W}$, и она представима в виде $\mathbf{W} = \sum_{i=1}^{p} w_i (U_i U'_i)$. Очевидно, здесь охвачены случаи $\mathbf{W} = \mathbf{I}$ и $\mathbf{W} = \mathbf{S}$. После несложных преобразований получаем следующую формулу для функционала качества (8.26):

$$L_W^2 = \mathbf{E} \sum_{i=1}^{p} w_i (\tilde{g}_i - g_i)^2. \qquad (8.48)$$

Чтобы получить аналитическое выражение L^2_W, запишем его в виде

$$L^2_W = \mathbf{E} \sum_{i=1}^{p} w_i \left(\frac{c_i}{\lambda_i} - \frac{a_i}{\lambda_i} \widehat{c_i} \right)^2, \quad (8.49)$$

где

$$c_i = C'_{yX} U_i, \widehat{c_i} = \widehat{C}'_{yX} U_i.$$

Взяв теперь математическое ожидание, получим

$$L^2(\widetilde{G}, G) = \sum_{i=1}^{p} w_i \frac{c_i^2}{\lambda_i^2} ((a_i - 1)^2 + \lambda_i a_i^2 \sigma^2/n), \quad (8.50)$$

или в эквивалентной форме

$$L^2(\widetilde{G}, G) = \sum_{i=1}^{p} w_i \left[g_i^2 (a_i - 1)^2 + \frac{1}{\lambda_i} a_i^2 \sigma^2/n \right]. \quad (8.50')$$

Для дальнейшего анализа понадобится еще преобразованное выражение для нормированной суммы квадратов отклонений $\widehat{\Delta_n}$ $(\widetilde{G})$. Из (8.23) имеем

$$\widehat{\Delta_n}(\widetilde{G}) = (n - p - 1) \widehat{\sigma_y^2} (1 - \widehat{R^2_{y.X}}) + \sum_{i=1}^{p} \lambda_i (a_i - 1)^2 \widehat{g_i^2}. \quad (8.51)$$

Первое слагаемое в (8.51) соответствует применению мнк-оценки, а второе возникает, если хотя бы один из вкладов $a_i \neq 1$.

Укажем некоторые часто используемые типы оценок, представимые в виде (8.47).

Однопараметрическая гребневая регрессия [208, 209]. Стандартная запись этой оценки имеет вид[1]

$$\widetilde{\Theta}(k) = (\mathbf{S} + k\mathbf{I})^{-1} C_{yX}, \quad (8.52)$$

или, что более предпочтительно, когда диагональные элементы матрицы S различны,

$$\widetilde{\Theta} = (\mathbf{S} + k\mathbf{DG}(\mathbf{S}))^{-1} C_{yX}, \quad (8.52')$$

где $\mathbf{DG}$ (S) — диагональная матрица diag $(s_{11}, \ldots, s_{pp})$, $k > 0$ — малое число, так называемый параметр гребня.

[1] В литературе часто употребляется запись вида $(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}'Y$, которая для центрированных переменных эквивалентна (8.52).

От оценки вида (8.52′) можно перейти к оценке вида (8.52) с помощью нормировки матрицы S к матрице корреляций. Дальше будем рассматривать только оценки вида (8.51).

Собственные векторы U_i $(i = \overline{1, p})$ матрицы S являются и собственными векторами матрицы $S + kI$ с собственными числами $\mu_i = \lambda_i + k$. Следовательно, матрица $(S + kI)^{-1} = \sum_{i=1}^{p} \frac{U_i U_i'}{\lambda_i + k}$ и с учетом (8.48) и вида $\widehat{g_i}$ получаем, что относительные весовые коэффициенты a_i для оценки гребневой регрессии равны;

$$a_i = \frac{\lambda_i}{\lambda_i + k}. \qquad (8.52'')$$

Значение параметра k подбирается из решения минимизационной задачи для функционала (8.50).

Многопараметрическая гребневая регрессия [192]. Стандартная запись соответствующей оценки имеет вид.

$$\tilde{\theta}_{гр}(k) = (S + K)^{-1} C_{yx}, \qquad (8.53)$$

где K — матрица, перестановочная с S. Собственные числа этой матрицы пусть будут $k_1, \ldots, k_p$. После несложного пересчета получаем, что веса вкладов главных компонент для этой модели равны:

$$a_i = \frac{\lambda_i}{\lambda_i + k_i} \quad (i = \overline{1, p}). \qquad (8{,}53)$$

Значения параметров k_i $(i = \overline{1, p})$ подбираются из решения оптимизационной задачи для функционала (8.50).

Оценка Марквардта [227] (оценка дробного ранга). Для этой оценки определяются два параметра: ранг r и вес α. Веса a_i имеют вид

$$a_i = \begin{cases} 1, & i < r; \\ 0 \leqslant \alpha \leqslant 1, & i = r; \\ 0, & i > r. \end{cases}$$

Методы определения ранга r и α приведены в [227].

Регрессия на главные компоненты. Веса a_i могут принимать одно из двух значений: $a_i = 1$, если выполняется какое-либо из условий информативности данной главной компоненты (см. п. 8.2), либо $a_i = 0$, если данная компонента удаляется. Заметим, что редуцированные оценки Джеймса — Стейна и

Мейера — Уилке также могут быть легко представлены в терминах весовых коэффициентов a_i.

8.5.1. Оптимальное взвешивание вклада главных компонент. Найдем теперь значения вкладов a_i, минимизирующие функционал $L^2_W(\tilde{G}(A), G)$ (8.50). Для этого учтем, что функционал (8.50) представляет собой сумму квадратичных по a_i слагаемых, каждое из которых является функцией только одного параметра a_i и не зависит от весовых коэффициентов w_i в функционале качества.

Значения a_i, минимизирующие функцию потерь, будут определяться простыми выражениями

$$a_i = g_i^2 \Big/ \left(g_i^2 + \frac{1}{n}\sigma^2/\lambda_i\right) \qquad (8.54)$$

и *не зависят от весовой матрицы* **W**.

Минимальное значение величины $L^2_W(\tilde{G}(A^*), A)$, соответствующее точке минимума $A^* = (a_1^*, \ldots, a_p^*)$, будет равно:

$$L^2_W(\widehat{G}, G) = \sum_{i=1}^{p} w_i \frac{\sigma^2}{n\lambda_i} \cdot \frac{g_i^2}{g_i^2 + \frac{1}{n}\sigma^2/\lambda_i}, \qquad (8.55)$$

в то время как для мнк-оценки

$$L^2_W(\widehat{G}, G) = \sum_{i=1}^{p} w_i \frac{\sigma^2}{n\lambda_i}.$$

Оценка $\tilde{G}(A^*)$, соответствующая оптимальному значению A^*, обладает следующими свойствами:

1) средний квадрат отклонения любого коэффициента $\tilde{g}_i(A^*)$ при i-й главной компоненте от истинного значения g_i меньше, чем для мнк-оценки $\widehat{g}_i$. Действительно, в силу (8.49) каждый член суммы в (8.55) представляет собой средний квадрат отклонения коэффициента $\tilde{g}_i(A^*)$ от истинного значения g_i, т. е. $\mathbf{E}(\tilde{g}_i(A^*) - g_i)^2 = \frac{\sigma^2}{n\lambda_i} \cdot \frac{g_i^2}{g_i^2 + \sigma^2/\lambda_i n}$, что меньше соответствующей величины $\frac{\sigma^2}{n\lambda_i}$ для мнк-оценки;

2) среднеквадратическое отклонение любого из параметров $\tilde{\theta}_i$, $(i = \overline{1,p})$ оценки $\tilde{\Theta}(A^*)$ (8.48) для переменных $x^{(1)}, \ldots, x^{(p)}$ от истинного значения Θ меньше, чем у мнк-оценок для соответствующих параметров [192];

3) для применения выражения (8.54) важным является то, что в точке $a_i = 1$ ($i = \overline{1,p}$) первая производная нормированной суммы квадратов отклонений $\widehat{\Delta}_n$ ($\widetilde{G}$ (A)) по a_i равна 0 (см. формулу (8.51)), и, следовательно, величина $\widehat{\Delta}_n$ ($\widetilde{G}$ (A)) в окрестности точки $a_i = 1$ ($i = \overline{1,p}$) меняется медленно. В то же время первая производная L^2 ($\widetilde{G}$ (A), G) в окрестности точки $a_i = 1$ ($i = \overline{1,p}$) положительна. Это позволяет надеяться, что можно подобрать такие значения $a_i < 1$, что значение величины $\widehat{\Delta}_n$ ($\widetilde{G}$ (A)) возрастет ненамного, а значение функционала L^2 ($\widetilde{G}$ (A), G) при этом уменьшится достаточно заметно.

В заключение заметим, что многопараметрическая гребневая регрессия (8.53), основанная на определении значений параметров гребня k_i ($i = \overline{1,p}$), которые минимизируют функционал (8.49), полностью эквивалентна регрессии с оптимальными весами вкладов главных компонент.

8.5.2. Оценка оптимальных вкладов главных компонент. Возникает вопрос, как воспользоваться формулой (8.54) на практике, если обе величины g_i и σ^2 неизвестны? Следуя [207], рассмотрим два метода оценивания a_i.

1. Вместо значений g_i^2 подставляем в (8.54) мнк-оценки $\widehat{g}_i$, а в качестве оценки для σ^2 берем величину $s^2 = \frac{n\widehat{\Delta}_n(\widehat{\Theta})}{n-p-1}$. Тогда получим оценки

$$\widehat{a}_i^{(1)} = \widehat{g}_i^2 / (\widehat{g}_i^2 + s^2/\lambda_i n). \tag{8.56}$$

2. Можно организовать итеративную процедуру следующим образом:

$$\begin{cases} \widehat{a}_i^{(t+1)} = (\widehat{a}_i^{(t)} \widehat{g}_i)^2 / ((\widehat{a}_i^{(t)} \widehat{g}_i)^2 + s^2/n\lambda_i); \\ \widehat{g}_i^{(t)} = \widehat{a}_i^{(t)} \widehat{g}_i. \end{cases}$$

Значения $a_i^{(1)}$ получаем из (8.56). Величина s^2 остается неизменной на всех итерациях. Аналогично [207] можно показать, что такой итеративный процесс сходится. Предельное значение $\widetilde{a}_i$ должно удовлетворять уравнению $\widetilde{a} = (\widetilde{a}^2 \widehat{g}_i^2)/(\widetilde{a}^2 \widehat{g}_i^2 + s^2/n\lambda_i)$. Это уравнение имеет три корня: $a = 0$ и корни, удовлетворяющие квадратному уравнению $\widetilde{a}^2 \widehat{g}_i^2 - \widetilde{a}\widehat{g}_i^2 + s^2/$

$/n\lambda = 0$. Последнее имеет вещественные корни, когда выполняется условие

$$n\lambda_i \widehat{g}_i^2/s^2 \geqslant 4. \tag{8.57}$$

Отсюда получаем, что

$$\tilde{a}_i = \begin{cases} 0, \text{ если } n\lambda_i \widehat{g}_i^2/s^2 < 4; \\ \frac{1}{2} + \frac{1}{\widehat{g}_i}\sqrt{\widehat{g}_i^2 - 4s^2/\lambda_i n}, \text{ если } n\lambda_i \widehat{g}_i^2/s^2 \geqslant 4. \end{cases} \tag{8.58}$$

Заметим, что отношение $n\lambda_i\widehat{g}_i^2/s^2$ есть квадрат t-статистики, использованной в § 8.2 в одном из методов выделения существенных главных компонент. Таким образом, отношение (8.57) устанавливает еще одну границу для объявления коэффициентов при главных компонентах нулевыми.

8.6. Вопросы точности вычислительной реализации процедур линейного оценивания

8.6.1. Два метода получения мнк-оценок. Когда набор предсказывающих переменных и модель определены, мнк-оценки неизвестных параметров линейного уравнения регрессии можно определить путем решения одной из следующих четырех систем линейных уравнений:

1) $\mathbf{X}\Theta_p = Y$, (8.60)

где $\mathbf{X}$ — матрица данных, расширенная путем добавления строки из единиц; $\Theta_p' = (\theta_0, \Theta')$ — вектор размерности $(p+1)$, а θ_0 — свободный член уравнения регрессии;

2) $\mathbf{X}_c\Theta = Y_c$, (8.60′)

где $\mathbf{X}_c$ — центрированная матрица данных; Y_c — n-мерный вектор центрированных значений зависимой переменной y;

3) $(\mathbf{X}'\mathbf{X})\Theta_p = \mathbf{X}'Y$, (8.60″)

т. е. Θ_p является решением нормальной системы уравнений, связанной с системой (8.60);

4) $(\mathbf{X}_c'\mathbf{X}_c)\Theta = \mathbf{X}_c Y_c'$, (8.60‴)

т. е. Θ является решением нормальной системы уравнений, связанной с системой (8.60'').

Решение системы нормальных уравнений (8.60'') или (8.60''') начали применять для получения оценок коэффициентов регрессии раньше, чем непосредственное решение системы линейных уравнений (8.60). Последний метод стали использовать примерно с середины шестидесятых годов [193, 194] (см. также более поздние работы [142, 143]). Основанием для активной пропаганды непосредственного решения системы (8.60), минуя этап получения нормальных уравнений, является доказанная в [193] большая устойчивость численного решения уравнения (8.60) при наличии ошибок округления и представления данных в ЭВМ по сравнению с решением системы нормальных уравнений. Однако, как показано далее, увеличение устойчивости может быть обосновано лишь при некоторых предположениях относительно свойств системы уравнений (8.60), которые далеко не всегда имеют место на практике.

Вопрос о выборе способа численного решения имеет смысл лишь в том случае, когда погрешность вычисления оценок коэффициентов регрессии на ЭВМ сравнима по величине с их статистическим разбросом, который определяется формулой (8.8). Необходимым для этого условием, как мы увидим далее, является наличие мультиколлинеарности. Но при выраженной мультиколлинеарности с точки зрения статистической устойчивости оценок лучше переходить к решению регуляризованных (тем или иным способом) систем уравнений (8.60), (8.60'), (8.60''), (8.60'''). Для систем нормальных уравнений методами регуляризации будут уже рассмотренные метод главных компонент (см. § 8.2) и гребневая регрессия (см. § 8.5).

8.6.2. Оценки величин возмущений для решений центрированной и соответствующей ей нормальной системы уравнений. Пусть $\mathbf{A}'\Theta = C$ некоторая система линейных уравнений, матрица $\mathbf{A}'$ которой имеет размерность $q \times k$ (k не обязательно равно q), Θ — вектор размерности k, правая часть C — вектор размерности q.

Как показано в [39], решение такой системы, получаемое на ЭВМ[1], на самом деле совпадает с решением некоторой возмущенной системы уравнений

$$(\mathbf{A} + \mathbf{P})' \Theta^* = C + \Delta C.$$

---

[1]В том случае, когда система $A'\Theta = C$ не имеет решения в обычном смысле или имеет не единственное решение, под решением будет пониматься псевдорешение с минимальной нормой, т. е. так называемое нормальное псевдорешение [17].

Введем относительную величину возмущения решения Θ: $\delta\Theta = \|\Theta - \Theta^*\| / \|\Theta\|$.

Величина возмущения $\delta\Theta$ как функция возмущений ΔC,**P** зависит от двух характеристик системы уравнений:

1) числа обусловленности матрицы системы [39]

$$\varkappa(\mathbf{A}) = \rho_{max}(\mathbf{A}) / \rho_{min}(\mathbf{A}),$$

где ρ_{max}, ρ_{min} — соответственно наибольшее и наименьшее (ненулевое) сингулярные числа матрицы **A**. Если матрица **A** имеет ранг l, то у нее имеется l ненулевых сингулярных чисел $\rho_1 \geqslant \rho_2 \geqslant \ldots \geqslant \rho_l$ и $\rho_{max} = \rho_1$, $\rho_{min} = \rho_l$. Для сингулярных чисел матрицы **AA'** соответствующей нормальной системы уравнений имеют место равенства [39]

$$\rho_{max}(\mathbf{AA}') = \rho^2_{max}(\mathbf{A}),\ \rho_{min}(\mathbf{AA}') = \rho^2_{min}(\mathbf{A}),$$

поэтому

$$\varkappa(\mathbf{AA}') = \varkappa^2(\mathbf{A});$$

2) величины относительной несогласованности системы

$$\Delta^2_{отн}(\mathbf{A}, C) = \min_{\Theta} \|\mathbf{A}\Theta - C\|^2 / \|C\|^2.$$

Для согласованной системы уравнений $\Delta^2_{отн}(\mathbf{A}, C) = 0$.

Рассмотрим теперь для определенности центрированную систему уравнений (8.60″), т. е. $\mathbf{A} = \mathbf{X}_c$, $C = Y$, $q = n$, $k = p$. Тогда верно следующее утверждение.

Утверждение. *Квадрат величины относительной несогласованности для центрированной системы уравнений*

$$\Delta^2_{отн}(\mathbf{X}_c, Y) = \begin{cases} 0, \text{ когда } n < p+1; \\ (1 - \widehat{R}^2_{y.x}) \dfrac{n-p-1}{n}, & n \geqslant p+1. \end{cases}$$

Соответствующая нормальная система уравнений всегда согласована, поэтому $\Delta^2_{отн}(\mathbf{X}'_c\mathbf{X}_c, \mathbf{X}'_c Y) = 0$.

Используя результаты [39, п. 37; 257], запишем теперь следующие оценки сверху для относительных погрешностей решений центрированной системы (8.60″) и нормальной системы (8.60‴):

$$\frac{\|\Theta^* - \Theta\|}{\|\Theta\|} \leqslant 4{,}9\,(l\varkappa(\mathbf{X}_c) + l\,(\varkappa^2(\mathbf{X}_c) + 1)\,\Delta_{отн}(\mathbf{X}_c, Y))\,\varepsilon_{маш};$$

$$\frac{\| \Theta^*_{\text{норм}} - \Theta \|}{\| \Theta \|} \leqslant 3\varkappa^2 (\mathbf{X}_c)\, \varepsilon_{\text{маш}},$$

где $\varepsilon_{\text{маш}}$ — машинная ошибка округления [136, п.2.2].

Когда система плохо обусловлена, величина $\varkappa (\mathbf{X}_c)$ велика, и основную часть погрешности определяют слагаемые с множителем $\varkappa^2 (\mathbf{X})$.

Из приведенных неравенств можно видеть, что с точки зрения влияния квадрата числа обусловленности $\varkappa(\mathbf{X})$ на верхнюю границу погрешности решения решать систему (8.60) выгоднее, если система неопределенная ($n < p$), а в случае переопределенной системы полного ранга ($n > p$, $l = p$), только если выполняется неравенство $4{,}9\, p \sqrt{\frac{n-p}{n}} (1 - \widehat{R}^2_{y.x}) < 3$, т. е. если коэффициент множественной корреляции между y и $\mathbf{X}$ достаточно велик. Для нецентрированной системы получается аналогичный результат.

Сравнивая два способа решения систем (8.60) (непосредственно с матрицей $\mathbf{X}$ и с переходом к системе нормальных уравнений), можно сделать вывод, что *несогласованные системы* (8.60), *как правило, лучше решать, используя переход к нормальной системе уравнений.* В статистической практике несогласованные системы возникают, когда матрица данных $\mathbf{X}$ переопределена, т. е. число объектов (столбцов) в ней больше числа переменных (строк), и при этом линейные уравнения, входящие в систему (8.60), не могут выполняться точно. Но превышение числа объектов над числом переменных — типичная ситуация в регрессионном анализе. Второе условие несогласованности также часто выполняется, так как обычно системы линейных уравнений используются для оценки параметров линейных моделей типа (8.1), являющихся лишь приближением действительных соотношений между переменными (мерой этого приближения как раз и является дисперсия случайной компоненты ε). Для обоснования перехода к нормальной системе уравнений существенно и то, что матрица $\mathbf{X}'\mathbf{X}$ тесно связана с ковариационной матрицей, которая является исходным объектом для различных видов многомерного анализа (главных компонент, факторного анализа и т. д.).

8.6.3. Центрирование и нормирование матрицы данных. Рассмотрим более подробно, как связаны решения систем нормальных уравнений для центрированной $\mathbf{X}_c$ и расширенной матриц данных.

Так как элементы первой строки расширенной матрицы данных полагаются равными единице, система нормальных урав-

нений (8.60'') имеет вид

$$\begin{bmatrix} n & n\overline{x}^{(1)} & \ldots & n\overline{x}^{(p)} \\ n\overline{x}^{(1)} & \sum_{i=1}^{n} x_i^{(1)} x_i^{(1)} & \ldots & \sum_{i=1}^{n} x_i^{(1)} x_i^{(p)} \\ \cdot & \cdot & \cdot & \cdot \\ n\overline{x}^{(p)} & \sum_{i=1}^{n} x_i^{(1)} x_i^{(p)} & \ldots & \sum_{i=1}^{n} x_i^{(p)} x_i^{(p)} \end{bmatrix} \begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_p \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \sum_{i=1}^{n} x_i^{(1)} y_i \\ \sum_{i=1}^{n} x_i^{(p)} y_i \end{pmatrix}, \tag{8.61}$$

где $\overline{x}^{(j)}$ — среднеарифметическое значение переменной $x^{(j)}$, а суммирование в выражениях для элементов вектора $\mathbf{X}'Y$ и матрицы $\mathbf{X}'\mathbf{X}$ проводится от 1 до n.

Если решать систему (8.61) методом последовательного исключения Гаусса или приведением матрицы $\mathbf{X}'\mathbf{X}$ к треугольной форме, то первый шаг состоит в делении первого уравнения на n и вычитании соответствующих кратных первого уравнения (1-й строки матрицы $\mathbf{X}'\mathbf{X}$) из остальных уравнений (строк матрицы $\mathbf{X}'\mathbf{X}$) таким образом, чтобы оставшиеся p элементов первого столбца матрицы $\mathbf{X}'\mathbf{X}$ обратились в нуль. Таким образом после первого шага мы получим систему уравнений вида

$$\begin{bmatrix} 1 & \overline{x}^{(1)} \ldots \overline{x}^{(p)} \\ 0 & \\ \vdots & nS \\ 0 & \end{bmatrix} \begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_p \end{pmatrix} = \begin{pmatrix} \overline{y} \\ n\widehat{C}_{yX} \end{pmatrix}. \tag{8.61'}$$

При этом элементы ковариационной матрицы s_{jk} $(i,\ j = \overline{1,p})$ фактически вычисляются по формуле

$$ns_{jk} = \sum_{i=1}^{n} x_i^{(j)} x_i^{(k)} - \overline{x}^{(j)} \overline{x}^{(k)}. \tag{8.62}$$

Хотя выражение (8.62) теоретически эквивалентно выражению

$$ns_{jk} = \sum_{i=1}^{n} (x_i^{(j)} - \overline{x}^{(j)}) (x_i^{(k)} - \overline{x}^{(k)}), \tag{8.63}$$

однако при реализации на ЭВМ формула (8.63) позволяет вычислять элементы s_{jk} с существенно меньшей погрешностью (особенно когда n велико), чем формула (8.62) (подробнее см. в п. 8.6.4). Из первого уравнения системы (8.61′) следует, что

$$\widehat{\theta}_0 = \bar{y} - \sum_{j=1}^{p} \widehat{\theta}_j \bar{x}^{(j)}, \tag{8.64}$$

а вектор $\widehat{\Theta}' = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$ является решением системы $n\mathbf{S}\widehat{\Theta} = nC_{yx}$, т. е. системы (8.60‴), поскольку $n\mathbf{S} = \mathbf{X}_c'\mathbf{X}_c$.

Таким образом, решение нормальной системы уравнений для расширенной матрицы данных сводится к решению системы нормальных уравнений с центрированной матрицей данных не только теоретически, но и во многих случаях при практической реализации вычислительной процедуры. Отметим в связи с этим следующее.

1. Согласно теореме о разделении собственных чисел [102] имеют место неравенства

$$\lambda_{\max}(\mathbf{X}'\mathbf{X}) \geqslant \lambda_{\max}(\mathbf{X}_c'\mathbf{X}_c) \text{ и } \lambda_{\min}(\mathbf{X}'\mathbf{X}) \leqslant \lambda_{\min}(\mathbf{X}_c'\mathbf{X}_c),$$

где $\lambda_{\max}(\Sigma)$, $\lambda_{\min}(\Sigma)$ — соответственно максимальное и минимальное собственные числа матрицы Σ.

Поэтому для чисел обусловленности имеет место неравенство $\varkappa(\mathbf{X}_c'\mathbf{X}_c) \leqslant \varkappa(\mathbf{X}'\mathbf{X})$, т. е. центрированная система, как правило, лучше обусловлена, чем система с расширенной матрицей данных.

2. Вычисление элементов ковариационной матрицы $\mathbf{S}$ проводится по неудовлетворительной, при реализации на ЭВМ, формуле (8.62), что может привести к возникновению дополнительной погрешности в решении. Поэтому если переходить к системе нормальных уравнений, то целесообразнее получать устойчивую (в вычислительном отношении) оценку ковариационной матрицы $\mathbf{S}$ (см. п. 8.6.4), решать систему вида (8.60‴) или эквивалентную ей систему $\mathbf{S}\widehat{\Theta} = C_{yx}$, а значение свободного члена $\widehat{\theta}_0$ получать из (8.64).

Дальнейшее улучшение обусловленности системы (8.60‴) и повышение точности вычислительной процедуры можно получить, переходя к нормированным переменным [163].

8.6.4. Вычисление элементов ковариационной матрицы. Коэффициенты системы линейных уравнений для центрированных переменных являются элементами матрицы ковариаций с точностью до множителя n. В связи с этим возникает задача аккуратного вычисления элементов матрицы ковариаций, чтобы избежать внесения дополнительной погрешности в решение ис-

ходной системы (8.69′) при переходе к соответствующей нормальной системе уравнений. Для этого следует воспользоваться так называемой двухэтапной оценкой

$$a_{jl} = ns_{jl} = \sum_{i=1}^{n} \left(x_i^{(j)} - \overline{x}^{(j)}\right)\left(x_i^{(l)} - \overline{x}^{(l)}\right). \tag{8.65}$$

Эта оценка названа двухэтапной, поскольку требует предварительного вычисления средних значений $\overline{x}^{(j)}$. Довольно часто в литературе по регрессионному анализу предлагается использовать оценку вида

$$a_{jl} = \sum_{i=1}^{n} x_i^{(j)} x_i^{(l)} - \frac{1}{n} \overline{x}^{(j)} \overline{x}^{(l)}. \tag{8.66}$$

Эта оценка обладает определенным преимуществом перед двухэтапной оценкой (8.65) с точки зрения организации вычислений, поскольку позволяет вычислить элементы a_{jl} за один просмотр данных. Однако она является неудовлетворительной в отношении величины погрешности, с которой вычисляются элементы ковариационной матрицы.

Приведем некоторые результаты, позволяющие сравнить точность оценки диагональных элементов a_{ij} при использовании формул (8.65) и (8.66). Далее для упрощения формул опустим индекс номера переменной и будем считать, что оценивается дисперсия некоторой переменной x соответственно по одной из двух схем:

$$a_1^2 = ns_1^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2; \tag{8.67}$$

$$a_2^2 = ns_2^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \overline{x}^2. \tag{8.67′}$$

Теоретически $s_1^2 = s_2^2 = s_x^2$. Для оценки погрешности введем, следуя [173], *число обусловленности данных* $k = ||x||/\sqrt{n}s_x$, где s_x^2 — точное значение дисперсии x, $||x||^2 = \sum_{i=1}^{n} x_i^2$.

Легко видеть, что значение $k \geqslant 1$ и оно возрастает, когда s_x^2 убывает при фиксированном значении $||x||$. Для относительной погрешности оценок (8.67) и (8.61′) верны следующие неравенства:

$$\left|\frac{s_2^2 - s_x^2}{s_x^2}\right| < 3nk^2 \varepsilon_{\text{маш}}; \quad \left|\frac{s_1^2 - s_x^2}{s_x^2}\right| < n\varepsilon_{\text{маш}} + n^2 k^2 \varepsilon_{\text{маш}}^2,$$

где $\varepsilon_{маш}$ — машинная ошибка округления.

Для реальных задач $n\varepsilon_{маш} \ll 1$ и, следовательно, двухэтапная оценка существенно точнее оценки (8.66), особенно когда значение числа обусловленности для данных k велико. В некоторых случаях оценка (8.66) может дать даже отрицательные значения для s_x^2. Не вдаваясь в детальный анализ, можно сказать, что относительно низкая точность оценки (8.66) объясняется тем, что она представляет собой разность двух неотрицательных величин, которые при больших k (малых значениях s_x^2) близки друг другу. При вычислении на ЭВМ такая ситуация как раз и приводит к потере точности.

В некоторых ситуациях, например, когда объем данных велик, и они размещены во внешней памяти, желательно избежать двукратного считывания данных при вычислении элементов ковариационной матрицы. Для этого можно использовать оценки типа скользящего среднего, которые позволяют вычислять ковариационную матрицу с той же относительной погрешностью, что и двухэтапная оценка. Приведем один из возможных вариантов алгоритмов вычисления элементов a_{jl} [259]:

$$a_{jl}^{(0)} = 0,\ t_{(0)}^{j} = 0,\ t_{(0)}^{l} = 0;$$

$$\left.\begin{aligned} t_{(i)}^{j} &= t_{(i-1)}^{j} + x_i^{(j)}, \\ t_{(i)}^{l} &= t_{(i-1)}^{(l)} + x_i^{(l)}, \end{aligned}\right\} \text{ для } i = \overline{1, n}; \tag{8.68}$$

$$a_{jl}^{(i)} = a_{jl}^{(i-1)} + \left(ix_i^{(j)} - t_{(i)}^{j}\right)\left(ix_i^{(l)} - t_{(i)}^{l}\right)/i\,(i-1)\ \left(\text{для } i = \overline{2, n}\right).$$

Однако этот алгоритм без дополнительных затрат памяти нельзя использовать при наличии пропущенных наблюдений.

Когда значения k и n велики, величина погрешности для двухэтапного алгоритма может стать недопустимо большой. Один из простых способов улучшения оценки в этом случае состоит в вычислении средних значений $\overline{x}^{(j)}$ с двойной точностью. Тогда имеет место следующее неравенство для погрешности ошибки:

$$\left|\frac{s_1^2 - s^2}{s^2}\right| < n\varepsilon_{маш} + O\left(\varepsilon_{маш}^3\right).$$

Для оценки скользящего среднего этого же эффекта можно добиться, накапливая с двойной точностью значения t^j и t^l.

О некоторых дальнейших возможностях повышения точности оценок ковариационной матрицы см. [173].

8.7. Отбор существенных переменных в задачах линейной регрессии

8.7.1. Влияние отбора переменных на оценку уравнения регрессии. Один из подходов к оцениванию параметров уравнения регрессии при наличии мультиколлинеарности состоит в сокращении количества входящих в модель предсказывающих переменных путем отбора подмножества предсказывающих переменных, существенных для прогноза значений переменной y. Каким бы способом ни проводился отбор переменных, число обусловленности уменьшается с уменьшением числа регрессоров. Процедура отбора существенных переменных, рассматриваемая как процедура выбора модели, полезна и когда исходная матрица $\mathbf{X}'\mathbf{X}$ хорошо обусловлена. Но особенно она эффективна в условиях мультиколлинеарности, когда объясняющие переменные сильно коррелированы. Так, если две какие-либо переменные сильно коррелированы с y и друг с другом, то часто бывает достаточно включения в модель одной из них, а дополнительным вкладом от включения другой можно пренебречь.

Отбор существенных переменных в пространстве главных компонент рассмотрен в п. 8.3. Как там показано, он приводит к следующим результатам: с одной стороны, к некоторому увеличению наблюдаемого значения нормированной суммы квадратов отклонений $\widehat{\Delta_n}$, но одновременно к уменьшению среднеквадратического отклонения от соответствующих истинных значений параметров и к уменьшению средней ошибки прогноза для векторов X^*, не входящих в матрицу плана $\mathbf{X}$ (т. е. в обучающую выборку, см. п.11.3). Последнего можно достичь и при отборе существенных переменных в исходном пространстве (опять-таки за счет увеличения нормированной суммы квадратов отклонений на обучающей выборке). Фактически отбор переменных означает, что исходное множество из p переменных делится на два подмножества $X(p-q)$ и $X(q)$, состоящих из таких $p-q$ и q переменных, что коэффициенты регрессии при $p-q$ переменных, входящих в первое подмножество, полагаются равными нулю, а коэффициенты при q переменных из второго подмножества оцениваются по мнк (по окончании процедуры отбора для оценки можно использовать и методы, изложенные в § 8.2—8.5).

В предположении, что матрица данных $\mathbf{X}$ является неслучайной, возможны две точки зрения на оценку уравнения регрессии, полученную после отбора существенных предсказывающих переменных.

Первая точка зрения исходит из того, что модель регрессии (8.1) является истинной, и несмещенная оценка коэффициентов регрессии получается мнк путем решения системы уравнений (8.3) (в условиях мультиколлинеарности эта оценка может быть неудовлетворительной, но тем не менее несмещенной). Тогда принудительное приравнивание части коэффициентов регрессионного уравнения к 0, что и происходит при отборе переменных, естественно, приводит, если матрица S недиагональна, к смещенным оценкам коэффициентов при оставшихся переменных, т. е. мы приходим к классу смещенных оценок, рассмотренных в § 8.3.

С другой стороны, процесс отбора существенных переменных можно рассматривать как процесс *выбора истинной модели* из множества возможных линейных моделей, которые могут быть построены с помощью набора предсказывающих переменных, и тогда полученные после отбора оценки коэффициентов можно рассматривать как несмещенные, хотя сама процедура отбора вводит некоторое смещение [93]. Этой точки зрения мы будем придерживаться далее.

Для случая, когда переменные $x^{(1)},\ldots, x^{(p)}$, y — случайные величины, вопрос о правильности (истинности) модели не возникает. Все, что мы ищем в этом случае,— это модель, сохраняющую ошибку предсказания на разумном уровне, при ограниченном количестве переменных.

8.7.2. Критерии качества уравнения регрессии. Любой алгоритм отбора существенных регрессоров выполняет следующую последовательность действий:

генерацию подмножеств переменных;

сравнение этих подмножеств по некоторому критерию качества уравнения регрессии, построенного по этим переменным;

проверку конца генерации (остановки алгоритма).

Рассмотрим наиболее употребительные критерии качества уравнения регрессии. Почти все они основаны на измерении средней величины ошибки прогноза, на векторах X, не вошедших в обучающую выборку (матрицу данных X), при тех или иных предположениях о распределении или способе формирования этих векторов.

1. **Коэффициент детерминации** (квадрат коэффициента множественной корреляции)

$$\widehat{R}^2_{y\,X} = 1 - \frac{n\widehat{\Delta}_n}{\sum_{i=1}^{n}(y_i-\bar{y})^2}.$$

Максимизация $\widehat{R}^2_{y\cdot X}$ эквивалентна минимизации нормированной остаточной суммы квадратов $\widehat{\Delta}_n$. В этом смысле $\widehat{R}^2_{y\cdot X}$ можно рассматривать как меру согласия модели с данными. Однако, поскольку в выражение для $\widehat{R}^2_{y\cdot X}$ входит и дисперсия переменной y, при анализе двух различных совокупностей данных (матриц (**X**, **Y**)) может иметь место ситуация, когда одна из регрессий имеет меньшее значение $\widehat{\Delta}_n$ и в то же время меньшее значение $\widehat{R}^2_{y\cdot X}$ за счет увеличения дисперсии σ^2_y. В случаях задачи отбора переменных это обстоятельство можно не учитывать, поскольку матрица данных не меняется и $R^2_{y\cdot X}$ можно рассматривать как относительную меру качества уравнения регрессии.

Недостаток $\widehat{R}^2_{y\cdot X}$ как критерия качества уравнения регрессии состоит в том, что значение коэффициента детерминации не убывает (по крайней мере) с ростом числа предсказывающих переменных, входящих в модель. Таким образом, модели, в которых больше переменных, будут более предпочтительными, если для сравнения использовать $\widehat{R}^2_{y\cdot X}$. Однако для сравнения уравнений регрессии с одинаковым числом зависимых переменных величина $\widehat{R}^2_{y\cdot X}$ является вполне подходящей. Некоторые из перечисленных ниже критериев являются монотонными функциями от $\widehat{R}^2_{y\cdot X}$, которые в то же время зависят от числа включенных в модель регрессоров q и объема выборки n и могут убывать с ростом $\widehat{R}^2_{y\cdot X}$.

2. **Скорректированный коэффициент детерминации.** Чтобы ввести скорректированный коэффициент детерминации, вспомним, что при $n \to \infty$ имеет место равенство $\sigma^2 = \sigma^2_y (1 - R^2_{y\cdot X(q)})$ или $R^2_{y\cdot X(q)} = 1 - \sigma^2/\sigma^2_y$. Для конечного объема обучающей выборки несмещенной оценкой для σ^2 является величина $s^2 = \widehat{\Delta}_n /(n - q - 1)$ (q — число регрессоров в модели), а для σ^2_y — величина

$$\widehat{\sigma}^2_y = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1).$$

Определим теперь скорректированный коэффициент детерминации из равенства $\widetilde{R}^2_{y\cdot X(q)} = 1 - s^2/\widehat{\sigma}^2_y$. После несложных преобразований получаем связь между обычным и скорректированным коэффициентами детерминации:

$$\widetilde{R}^2_{y\, X(q)} = 1 - \frac{n-1}{n-q-1}\left(1 - \widehat{R}^2_{y\cdot X(q)}\right). \quad (8.69)$$

В отличие от обычного скорректированный коэффициент дерерминации *может уменьшаться с ростом числа предсказывающих переменных* q, если в результате введения дополнительной переменной изменение $1 - \widehat{R}^2_{y \cdot X(q)}$ оказывается недостаточным для компенсации увеличения отношения $(n-1)/(n-q-1)$.

В отличие от обычного коэффициента детерминации скорректированный уменьшается с ростом числа предсказывающих переменных q, если в результате введения дополнительной переменной изменение $1 - \widehat{R}^2_{y \cdot X(q)}$ оказывается недостаточным для компенсации увеличения отношения $(n-1)/(n-q-1)$.

3. Статистика Мэллоуза C_q. В [225] предложено использовать так называемую C_q статистику как меру качества уравнения регрессии с q предсказывающими переменными. В принятых здесь обозначениях

$$C_q = \frac{(n-q-1)\,(1-R^2_{y.X(q)})}{1-R^2_{y.X(p)}} - n + 2q + 2. \qquad (8.70)$$

4. Средний квадрат ошибки предсказания СКОП. Этот критерий предлагается в [24] (см. также [164, 42, 52]). При введении этого критерия предполагается, что переменные $(y, x^{(1)}, \ldots, x^{(p)})$ являются случайными величинами и имеют в совокупности $(p+1)$-мерное распределение. Таким образом, матрица данных $(\mathbf{X}, \mathbf{Y})$ представляет собой выборку объема n из $(p+1)$-мерного нормального распределения.

Пусть теперь $\widehat{y}_{(q)}(X) = \bar{y} + \Theta'(q)\,(X(q) - \bar{X}(q))$ — функция регрессии, основанная на q из p возможных предсказывающих переменных, и $\widehat{\Theta}(q)$ — мнк-оценка вектора регрессионных коэффициентов для набора из q переменных, $\bar{X}(q)$ — q-мерный вектор средних значений для переменных $x^{(j)}$, принадлежащих набору $X(q)$. Пусть теперь уравнение регрессии используется для предсказания значения переменной y для некоторого нового случайного вектора X^*.

Величина СКОП определяется как

$$\text{СКОП}_{(q)} = \mathbf{E}\,(\widehat{y}_{(q)}(X^*) - y^*)^2,$$

где математическое ожидание берется по всем случайным переменным, в том числе и по «новому» наблюдению X^*. Если использовать понятия обучающей и контрольной выборки, то можно сказать, что СКОП определяет среднюю квадратическую ошибку прогноза на контрольной выборке.

В [251] показано, что

$$\text{СКОП}_{(q)} = K(n, q)\,\sigma^2_{y.X(q)},$$

где $K(n, q) = (n^2 - n - 2)/(n(n - q - 2))$, $n > q + 2$ и $\sigma_{y \cdot X(q)}$ — условная дисперсия y относительно q переменных, входящих в уравнение регрессии. При применении этого критерия неизвестное значение дисперсии $\sigma^2_{y\ X(q)}$ заменяется ее оценкой максимального правдоподобия:

$$\widehat{\sigma}^2_{y.X(q)} = (n - 1)\,s^2_y\,(1 - \widehat{R}^2_{y.X(q)})/(n - q - 1).$$

Окончательно используемая как критерий оценка имеет вид

$$\widehat{\text{СКОП}}_{(q)} = \frac{(n^2 - n - 2)(n - 1)\,s^2_y\left(1 - \widehat{R}^2_{y.X(q)}\right)}{n(n - q - 1)(n - q - 2)}. \qquad (8.71)$$

5. **Несмещенная оценка коэффициента множественной корреляции.** Если переменные $(y, x^{(1)}, \ldots, x^{(p)})$ имеют в совокупности многомерное нормальное распределение, то оценка квадрата коэффициента множественной корреляции $\widehat{R}^2_{y \cdot X(q)}$ является смещенной. Несмещенная оценка (с точностью до членов $O(1/n^2)$) определяется с помощью выражения

$$\widetilde{R}^2_{y.X(q)} = \widehat{R}^2_{y.X(q)} - \frac{(q - 2)\left(1 - \widehat{R}^2_{y.X(q)}\right)}{n - q - 1} - \frac{2(n - 3)\left(1 - \widehat{R}^2_{y.X(q)}\right)^2}{(n - q - 1)(n - q + 3)}. \qquad (8.72)$$

Эта величина также может быть использована как критерий качества уравнения регрессии.

8.7.3. Схемы генерации наборов переменных. Когда критерий качества набора предсказывающих переменных фиксирован для выбора оптимального или хотя бы «хорошего» набора, необходимо провести сравнение достаточно большого числа различных наборов переменных и выбрать среди них наилучший. Рассмотрим некоторые схемы генерации наборов, применяющиеся в настоящее время.

Схемы полного перебора («всех возможных регрессий», метод «ветвей и границ»). Задачу полного перебора можно сформулировать следующим образом: для $q = 1, \ldots, p - 1$ найти набор из q предсказывающих переменных с минимальным значением остаточной суммы квадратов $\Delta^2_{X(q)}$ или, что эквивалентно, с максимальным значением коэффициента детерминации $R^2_{y \cdot X(q)}$. Так как критерии, приведенные в

п. 8.7.2, являются монотонными функциями от $R^2_{y\cdot X(q)}$; то этот набор будет оптимальным и по любому из них. Число различных подмножеств из q переменных, если всего имеется p переменных, будет равно C^q_p (числу сочетаний по q элементов из p возможных), а полное число наборов при изменении q от 1 до p будет 2^p. Ясно, что это число очень быстро растет с ростом p. Так, при $p = 20$ оно будет примерно 10^6, а при $p = 30$ — 10^9. Все же на современных ЭВМ возможна реализация полного перебора для значений p порядка 15.

В связи с необходимостью просмотра большого числа регрессионных моделей особенно важное значение приобретает использование экономных (в смысле количества машинных операций) методов расчета значений критерия и коэффициентов для соответствующих регрессионных моделей. Поэтому процедура генерации последовательности наборов переменных должна удовлетворять двум требованиям. Во-первых, переход от набора к набору должен осуществляться путем добавления или отбрасывания только одной переменной, что позволяет использовать экономные схемы пересчета значений критерия (см. п. 8.7.4) вместо полного решения соответствующей новой задачи регрессии. Среднее число операций для прямого расчета регрессии с q переменными имеет порядок q^3, а формулы пересчета уменьшают среднее число операций до порядка q^2.

В [189] предложена еще более эффективная процедура пересчета, позволяющая сократить число операций до порядка q, если требуется вычисление коэффициентов регрессии, и 6, если вычисляется только величина $R^2_{y\cdot X(q)}$. При этом, однако, требуется дополнительная память для размещения p матриц размера $p \times p$.

Второе требование состоит в том, чтобы любой набор генерировался только один раз. Описания процедур генерации, удовлетворяющих этим требованиям, приведены в [189, 191, 248]. Если в качестве основного лимитирующего фактора принять время вычислений, то наилучшим из алгоритмов полного перебора в настоящее время следует признать алгоритм Фёрнивала, предложенный в [189].

Объем вычислений при прямом переборе с ростом p растет настолько быстро, что уже при $p \approx 20$ превышает реальные возможности большинства ЭВМ. Выход из положения ищут с помощью методов ветвей и границ. Смысл этого метода заключается во введении какого-либо грубого правила, которое позволяет отбросить большинство наборов, не вычисляя для них значения критерия в силу их бесперспективности. Такое правило может быть основано на неравенстве

$R^2_{y \cdot X(A)} \geqslant R^2_{y \cdot X(B)}$, где $X(A)$ — любой набор предсказывающих переменных, а $X(B)$ — его подмножество. Другими словами, при исключении из регрессии каких-либо переменных значение $\widehat{R}^2_{y \cdot X(A)}$ может только убывать. Пусть теперь мы знаем некоторую оценку снизу R^2_q для оптимального значения $\widehat{R}^2_{y \cdot X(q)}$. Если для какого-либо набора $X(j)$ $R^2_{y \cdot X(j)} < R^2_q$ и $j > q$, то, очевидно, все поднаборы размерности q, полученные из $X(j)$, являются бесперспективными и могут не рассматриваться.

Использование методов ветвей и границ позволяет рассматривать задачи с $p \approx 50$—70. Наиболее эффективной является реализация алгоритма, предложенная в работе Фёрнивала [190]. Подробное описание одного из алгоритмов, реализующего метод «ветвей и границ», — алгоритма Хокинга—Лесли [206] на русском языке приведено в [79].

8.7.4. Пошаговые процедуры генерации наборов. Существенного сокращения числа генерируемых для сравнения наборов предсказывающих переменных можно добиться с помощью пошаговых (STEP—WISE) процедур отбора переменных. Хотя ни одна из пошаговых процедур не гарантирует получения оптимального по заданному критерию набора переменных (соответствующие примеры приведены, например, в [226, 205, 79]), все же обычно получаемые с их помощью наборы переменных являются достаточно хорошими для практического применения. Кроме того, возможны простые модификации традиционных пошаговых схем, которые позволяют преодолеть ряд присущих им недостатков.

Основными пошаговыми процедурами генерации наборов являются *процедура последовательного присоединения*, *процедура присоединения-удаления* и *процедура последовательного удаления*.

Рассмотрим один из возможных способов организации вычислений в пошаговой процедуре последовательного присоединения.

На первом шаге из исходного набора предсказывающих переменных $X(p) = (x^{(1)}, \ldots, x^{(p)})$ выбирается переменная $x^{(j_1)}$, имеющая максимальное значение квадрата коэффициента парной корреляции с y, т. е.

$$j_1 = \arg\max_{1 < k \leqslant p} r^2_{yx(k)}.$$

Признак $x^{(j_1)}$ составляет информативный набор предсказывающих переменных $X(1)$. Применяя теперь к матрице **A**

прямой оператор симметричного выметания W_{j_1} (см. п. 8.7.5), получим матрицу $\mathbf{A}_1$ и переходим ко второму шагу.

Второй шаг состоит в следующем. Пусть уже построен информативный набор из q предсказывающих переменных $X(q) = (x^{(j_1)}, \ldots, x^{(j_q)})$; пусть $\mathbf{A}_q$ — матрица, полученная из исходной матрицы $\mathbf{A}$ путем применения оператора выметания по переменным из $X(q)$. Ищем переменную $x^{(j_{q+1})}$, имеющую максимальное значение квадрата коэффициента частной корреляции с y при фиксированных переменных из $X(q)$:

$$j_{q+1} = \underset{(j=1,\,p,\,x^{(k)} \notin X(q))}{\arg\max} \left(\widehat{r}^2_{yX(k)\,(X(q))}\right).$$

При этом как кандидаты на присоединение к набору $X(q)$ используются лишь переменные, для которых вычисляется условие (см. п. 8.7.5) $1 - R^2_{x^{(k)}\cdot X(q)} > \tau_{\text{пор}}$. Если таких переменных не окажется, то работа алгоритма (отбор переменных) прекращается.

Отбор переменной $x^{(j_{q+1})}$ из условия максимума квадрата частного коэффициента корреляции эквивалентен ее выбору из условия максимума коэффициента множественной корреляции между y и набором $X(q+1) = X(q) \oplus x^{(j_{q+1})}$, так как имеет место тождество (см., например, [24, п. 3.2.4])

$$1 - \widehat{R}^2_{y.X(q+1)} = (1 - \widehat{R}^2_{y.X(q)})\left(1 - \widehat{r}^2_{yx(j_{q+1})\,(X(q))}\right).$$

После определения переменной $x^{(j_{q+1})}$ проверяется условие остановки процедуры отбора.

Основные из используемых условий остановки следующие:

а) процедура останавливается, если отобрано заданное пользователем количество переменных k, т. е. если $q + 1 = k$. При этом переменная $x^{(j_{q+1})}$ присоединяется к набору $X(q)$, а к матрице $\mathbf{A}_q$ применяется оператор выметания по переменной $x^{(j_{q+1})}$;

б) проверяется гипотеза $H_0 : \widehat{r}^2_{q+1} = \widehat{r}^2_{yx(j_{q+1})\,(X(q))} = 0$, для чего вычисляется значение F-статистики

$$F_{q+1} = (n - q - 2)\,\widehat{r}^2_{q+1}/(1 - \widehat{r}^2_{q+1}).$$

Если величина $F_{q+1} < F_{\text{вкл}}$, где $F_{\text{вкл}}$ — некоторая заранее заданная величина, то переменная $x^{(j_{q+1})}$ не присоединяется к набору $X(q)$, который и считается результатом работы алгоритма.

Используемая статистика F_{q+1} формально совпадает со статистикой для проверки значимости соответствующего регрессионного коэффициента в обычной задаче регрессии. Поэтому в качестве значения для $F_{вкл}$, как правило, выбирают классические уровни значимости (5, 10, 15%), соответствующие F-распределению с 1 и $(n - q - 2)$ степенями свободы. Однако величина F_{q+1} в пошаговой процедуре на самом деле не подчиняется F-распределению с соответствующим числом степеней свободы, поскольку проверяется гипотеза о равенстве нулю *максимального* по абсолютной величине коэффициента частной корреляции из $p - q$ коэффициентов частной корреляции для переменных, не входящих в $X(q)$. Неизвестно поэтому, какому уровню значимости соответствует выбранное значение;

в) процедура останавливается, если достигнуто максимальное (минимальное) значение критерия качества набора переменных. Пусть K_q — текущее значение какого-либо из критериев п. 8.7.2. Тогда процедура останавливается, если выполняются условия $K_q < K_{q+1}$ для критериев (8.69), (8.72) или $K_q > K_{q+1}$ для критериев (8.70), (8.71). Результирующим считается набор $X(q)$.

Можно показать, что правило остановки по текущему значению критерия эквивалентно правилу остановки по значению F-статистики при некоторой величине $F_{вкл}$. О других способах использования критериев в правилах остановки см. в [164].

Если условие остановки не выполняется, то к матрице $\mathbf{A}_q$ применяется оператор прямого выметания по переменной $x^{(i_{q+1})}$, и путем включения в $X(q)$ переменной $x^{(i_{q+1})}$ формируется новый текущий информативный набор $X(q+1)$. Затем второй шаг повторяется для набора $X(q+1)$.

Пошаговая процедура последовательного присоединения-удаления переменных (обычно именуемая в литературе просто как процедура последовательного присоединения) была впервые предложена в [180]. Приводимое ниже описание процедуры имеет некоторые отличия от исходной процедуры Эфроимсона. Формирование информативного набора переменных в этой процедуре организовано следующим образом.

Первый шаг совпадает с первым шагом процедуры последовательного присоединения.

На втором шаге, начиная с $q = 3$, перед поиском присоединяемой переменной $x^{(i_{q+1})}$ добавляется подшаг поиска переменной $x^{(l)}$, которую целесообразно удалить из текущего набора $X(q)$. Для этого определяется переменная $x^{(l)} \in X(q)$,

удаление которой приводит к минимальному уменьшению коэффициента детерминации, т. е.

$$l = \arg\min_{x^{(k)} \in X(q)} \left(\widehat{R}^2_{y.X(q)} - \widehat{R}^2_{y.X_{-k}(q-1)}\right),$$

где $X_{-k}(q-1)$ — набор $X(q)$ с удаленной переменной $x^{(k)}$. После определения номера l целесообразность удаления переменной $x^{(l)}$ обычно проверяется на основе сравнения F-статистики для проверки гипотезы $H_0: R^2_{y \cdot X(q)} = R^2_{y}\, X_{-l(q-1)}$ или эквивалентной ей гипотезы о коэффициенте частной корреляции $H_0: r_{yx^{(l)}(X_{-l}(q-1))} = 0$ с некоторым заранее заданным пороговым значением $F_{искл}$. Обычно выбирают значение $F_{искл} > F_{вкл}$ (так чтобы исключить переменные из набора было труднее, чем добавлять) соответственно 2%-, 1%-, 0,5%-ному уровням значимости при F-распределении с 1 и $(n-q-2)$ степенями свободы. На самом деле по тем же причинам, что и при присоединении переменных, величина F-статистики при удалении переменных не подчиняется F-распределению, и точный уровень значимости неизвестен.

Другой способ определения целесообразности удаления переменной $x^{(l)}$ основан на проверке «улучшения» качества набора по какому-либо из критериев качества п. 8.7.2.

Если качество набора «улучшается», то переменная удаляется. При удалении переменной $x^{(l)}$ из $X(q)$ к матрице $\mathbf{A}_q$ применяется оператор обратного выметания U_l. После фазы удаления переменной проводится фаза расширения набора ($X(q)$, если не было удаления, и $X_{-l}(q-1)$, если была удалена переменная $x^{(l)}$), точно так же, как и в процедуре последовательного присоединения. Остановка процедуры присоединения-удаления проводится по тем же правилам, что и остановка процедуры последовательного присоединения.

Пошаговая процедура последовательного удаления (исключения). Перед началом работы процедуры необходимо получить матрицу $\mathbf{A}_p$. Именно она теперь является той исходной матрицей, к которой применяется последовательность операторов выметания W_k, U_k. Для этого необходимо вычислить $\mathbf{R}_X^{-1}$, $\Theta(X(p)) = \widehat{\Theta}$ и $R^2_{y \cdot X(p)}$.

Первый шаг процедуры последовательного удаления состоит в определении такой переменной $x^{(i_1)}$, удаление которой из исходного набора $X(p)$ приводит к минимальному увеличению остаточной суммы квадратов $\widehat{\Delta}_n$ или, что эквивалентно, к минимальному уменьшению коэффициента детерминации. Величина изменения коэффициента детерминации проверяется

на значимость таким же образом, как и в фазе удаления процедуры последовательного присоединения (q при этом заменяется на p). Можно также проверять «улучшение» качества набора по какому-либо из критериев. Если значение F-статистики превышает значение $F_{\text{удал}}$ или если произошло «улучшение» качества набора переменных, то формируется набор $X(p-1)$ с удаленной переменной $x^{(i_1)}$, а к матрице $\mathbf{A}_p$ применяется оператор обратного выметания U_{i_1}.

Второй шаг состоит в следующем. Пусть $X(q)$ — текущий информативный набор, полученный в результате удаления $(p-q)$ переменных, и $\mathbf{A}_q$ — матрица, полученная из $\mathbf{A}_p$ применением к ней $(p-q)$ операторов обратного выметания. В наборе $X(q)$ ищем переменную $x^{(i_q)}$, удаление которой из $X(q)$ приводит к минимальному уменьшению коэффициента множественной детерминации. Затем проверяется условие остановки. Могут быть использованы следующие условия остановки:

а) получение набора с заданным количеством k предикторных переменных, т. е. проверяется условие $k = q - 1$;

б) превышение порогового значения $F_{\text{искл}}$ величиной F-статистики для проверки гипотезы $H_0: R^2_{y \cdot X(q)} = R^2_{y \cdot X(q-1)}$;

в) отсутствие «улучшения» качества набора по какому-либо из критериев п. 8.7.2.

По поводу других правил остановки см. [24, п. 3.3.2].

Если выполняются условия остановки б) и в), информативным набором при выходе из процедуры считается набор $X(q)$, а при выполнении условия а) выходным будет набор $X(q-1)$, получаемый из $X(q)$ удалением переменной $x^{(i_q)}$, и к матрице $\mathbf{A}_q$ применяется оператор U_{i_q}.

Если остановки процедуры не происходит, то текущим информативным набором становится набор $X(q-1)$, к матрице $\mathbf{A}_q$ применяется оператор выметания U_{i_q}. После этого второй шаг повторяется в применении к набору $X(q-1)$.

Рассмотрим теперь один экономичный по количеству вычислений способ определения удаляемой переменной $x^{(l)}$. Он может быть использован и в фазе удаления переменной для процедуры присоединения-удаления.

Пусть $\widehat{\theta}_i(X(q))$ $(i = \overline{1, q})$ — оценка коэффициента уравнения регрессии y для переменной $x^{(j_i)} \in X(q)$. Эти коэффициенты являются соответствующими элементами матрицы $\mathbf{A}_q$, и, следовательно, проводить дополнительных вычислений не нужно. Предлагаемый метод расчета основан на следующем равенстве. Если из набора $X(q)$ удаляется переменная $x^{(j_k)}$,

то

$$\Delta R_k^2 = \widehat{R}^2_{y.X(q)} - \widehat{R}^2_{y.X_{-k}(q-1)} = \widehat{\theta}_k^2 / a^{j_k j_k},$$

где $a^{j_k j_k}$ — элемент обратной матрицы корреляции для переменной из $X(q)$; $X_{-k}(q-1)$ — набор переменных, полученный из $X(q)$ при удалении $x^{(j_k)}$. Значение величины $a^{j_k j_k}$ также может быть извлечено из матрицы $\mathbf{A}_q$. Напомним, что рассматриваемое равенство относится к нормированным переменным. Переменная, подлежащая удалению, определяется как

$$x^{(l)} = \arg\min_{x^{j_k} \in X(q)} \Delta R_k^2.$$

8.7.5. Оператор симметричного выметания. С вычислительной точки зрения пошаговые процедуры последовательного присоединения и присоединения-удаления удобно реализовать как последовательность операций выметания, примененных к исходной расширенной корреляционной матрице $\mathbf{A}$ размера $(p+1)(p+1)$, которую можно представить в виде следующей блочной матрицы

$$\mathbf{A} = \begin{bmatrix} \mathbf{R}_X & \mathbf{r}_{yX} \\ \mathbf{r}'_{yX} & 1 \end{bmatrix},$$

где $\mathbf{R}_X$ — матрица коэффициентов корреляции между предсказывающими переменными порядка $p \times p$; $\mathbf{r}_{yX}$ — p-мерный вектор коэффициентов корреляции независимой переменной y с предсказывающими переменными. Таким образом, при отборе переменных мы фактически переходим к нормированным предсказывающим переменным и y.

Рассматриваемый ниже оператор симметричного выметания предложен в [162, 191, 119 п. 12.2]. Будем различать оператор прямого выметания W_k по переменной $x^{(k)}$ (это соответствует расширению текущего набора за счет включения переменной $x^{(k)}$) и оператор обратного выметания U_k по переменной $x^{(k)}$ (что соответствует удалению переменной $x^{(k)}$ из текущего набора). Действие оператора выметания на матрицу $\mathbf{A}$ состоит в пересчете ее элементов по одной из следующих схем:

для оператора прямого выметания

$$W_k(\mathbf{A}) = \begin{cases} a_{kk}^{\text{нов}} = -1/a_{kk}; \\ a_{ik}^{\text{нов}} = a_{ki}^{\text{нов}} = a_{ik}\, a_{kk}^{\text{нов}} \quad (i = \overline{1, p+1};\ i \neq k); \\ a_{ij}^{\text{нов}} = a_{ji}^{\text{нов}} = a_{ij} + a_{kj}^{\text{нов}} \quad (i,j = \overline{1, p+1};\ i,j \neq k); \end{cases}$$

для оператора обратного выметания

$$U_k(\mathbf{A}) = \begin{cases} a_{kk}^{\text{нов}} = -1/a_{kk}; \\ a_{ik}^{\text{нов}} = a_{ki}^{\text{нов}} = -a_{ik}\,a_{kk}^{\text{нов}} \ (i = \overline{1, p+1};\ i \neq k); \\ a_{ij}^{\text{нов}} = a_{ji}^{\text{нов}} = a_{ij} - a_{ik}\,a_{kj}^{\text{нов}} \ (i, j = \overline{1, p+1};\ i, j \neq k). \end{cases}$$

Операторы выметания W_k, U_k обладают следующими важными свойствами:

а) обратимость

$$W_k(U_k(\mathbf{A})) = U_k(W_k(\mathbf{A})) = \mathbf{A};$$

б) коммутативность

$$W_k(W_l(\mathbf{A})) = W_l(W_k(\mathbf{A}));$$
$$U_k(U_l(\mathbf{A})) = U_l(U_k(\mathbf{A})).$$

Эти свойства легко интерпретируются в терминах включения и исключения переменных $x^{(k)}$ и $x^{(l)}$ в текущий набор;

в) оба оператора сохраняют симметрию матрицы $\mathbf{A}$. Благодаря свойству в) при вычислениях необходимо использовать только верхний треугольник матрицы $\mathbf{A}$, что позволяет вдвое сократить необходимую память и объем вычислений.

Предположим, что в результате работы какой-либо процедуры отбора получен информативный набор $X(q)$ из q предсказывающих переменных и при этом применялся пересчет элементов матрицы $\mathbf{A}$ с помощью соответствующей последовательности операторов выметания W_k, U_k. Для упрощения обозначений будем считать, что в набор $X(q)$ включены q первых переменных $x^{(1)}, \ldots, x^{(q)}$ (этого всегда можно добиться перенумерацией переменных из X). Тогда результирующая матрица $\mathbf{A}_q$ будет иметь следующую структуру:

$$\mathbf{A}_q = \left[\begin{array}{c|c|c} -\mathbf{R}^{-1}_{X(q)} & -\mathbf{B}(X(q)) & -\widehat{\Theta}(X(q)) \\ \hline -\mathbf{B}'(X(q)) & \mathbf{C}_{X(p-q)(X(q))} & C_{yX(p-q)(X(q))} \\ \hline -\widehat{\Theta}'(X(q)) & C'_{yX(p-q)(X(q))} & 1 - \widehat{R}^2_{y.X(q)} \end{array}\right],$$

где $\mathbf{R}^{-1}_{X(q)}$ — матрица размера $q \times q$, обратная к матрице корреляций переменных из $X(q)$; $\mathbf{C}_{X(p-q)(X(q))}$ — матрица размера $(p-q) \times (p-q)$ частных ковариаций нормированных переменных $X(p-q) = (x^{(q+1)}, \ldots, x^{(p)})$, не включенных в

информативный набор; $\mathbf{B}(X(q))$ — матрица размера $q \times (p-q)$, компоненты i-го столбца которой представляют собой коэффициенты регрессии нормированной переменной $x^{(q+i)} \in X(p-q)$ на нормированные переменные из $X(q)$; $\widehat{\Theta}(X(q))$ — q-мерный вектор коэффициентов регрессии нормированной переменной y на нормированные переменные из $X(q)$; $\widehat{C}_{yX(p-q)(X(q))}$ — $(p-q)$-мерный вектор частных коэффициентов ковариаций нормированных переменных из $X(p-q)$ с y; $\widehat{R}^2_{y\cdot X(q))}$— квадрат коэффициента множественной корреляции между переменной y и предсказывающими переменными из $X(q)$.

Таким образом, матрица $\mathbf{A}_q$ содержит полное решение задачи регрессии независимой переменной y на переменные из $X(q)$ за исключением значения свободного члена. Из нее также легко извлечь частные коэффициенты корреляции переменных $r_{yx^{(q+i)}(X(q))}$ с y, необходимые для продолжения пошаговых процедур. Именно

$$r_{yx^{(q+i)}(X(q))} = c^{(i)}_{yX(p-q)(X(q))} \Big/ \left(\sqrt{1 - c^{ii}_{X(p-q)(X(q))}} \sqrt{1 - R^2_{y\cdot X(q)}} \right),$$

где $c^{(i)}_{yX(p-q)X(q)}$ — i-й элемент вектора частных ковариаций; $c^{ii}_{X(p-q)(X(q))}$ — диагональный элемент (остаточная дисперсия нормированной переменной $x^{(q+i)}$) матрицы частных ковариаций $\mathbf{C}_{X(p-q)(X(q))}$.

Зная значение $\widehat{R}^2_{y\cdot X(q)}$, легко вычислить и значения критериев качества уравнения регрессии, приведенных в п. 8.7.2.

Диагональные элементы матрицы частных ковариаций $\mathbf{C}_{X(p-q)(X(q))}$ представляют собой остаточные дисперсии нормированных переменных $x^{(q+i)} \in X(p-q)$ относительно переменных из $X(q)$ и могут быть записаны в виде

$$\tau_i = c^{ii}_{X(p-q)(X(q))} = 1 - R^2_{x^{(q+i)}\cdot X(q)}.$$

В условиях мультиколлинеарности значения $R^2_{x^{(q+i)}\cdot X(q)}$ для некоторых переменных $x^{(q+i)}$ могут быть очень близки к 1. При попытке добавить такую переменную в информативный набор необходимо использовать величину, обратную к τ_i, что при чрезмерной малости последней может привести к вычислительным трудностям. Поэтому целесообразно ввести пороговое значение, которое запретило бы использовать переменную $x^{(q+i)}$, если соответствующее значение τ_i будет мень-

ше порогового, т. е. если выполнится неравенство $\tau_i < \tau_{пор}$, то переменная $x^{(q+i)}$ не будет использоваться для расширения набора $X(q)$. Если же это неравенство выполняется для всех переменных из $X(p-q)$, то отбор переменных следует считать оконченным.

8.7.6. Методические аспекты использования процедур отбора существенных предикторных переменных. Когда число потенциальных переменных велико, формальное применение любой из рассмотренных процедур отбора может привести к неудовлетворительному с содержательной точки зрения набору предикторных переменных.

Рассмотрим некоторые методические приемы, позволяющие увеличить эффективность применения пошаговых процедур отбора.

1. Повторное применение процедур отбора с принудительно включаемыми переменными (ПВП). Возможность принудительного (обязательного) включения переменных в выходной набор $X(q)$ достигается достаточно простой модификацией описанных пошаговых процедур, а также методов «всех регрессий» и «ветвей и границ».

При использовании ПВП в процедурах последовательного присоединения и присоединения-удаления формирование выходного информативного набора происходит путем расширения начального набора, состоящего из ПВП, а для процедуры последовательного удаления переменная, удаляемая на каком-либо шаге, не должна входить в число ПВП.

Если имеется возможность использовать ПВП, целесообразно провести, помимо автоматизированного отбора, также и несколько вариантов отбора с различными ПВП. Окончательный набор получится в результате сравнения найденных наборов. Состав ПВП определяется, например, из экспертных соображений. Другой возможный подход к формированию ПВП основан на анализе графика какого-либо из критериев качества набора, выводимого при работе пошаговых процедур. С этой целью отбор переменных целесообразно проводить по возможности до исчерпания всего исходного множества потенциальных переменных с одновременным выводом на каждом шаге значений коэффициентов детерминации и критериев качества набора. Такой режим легко осуществить, если в процедуре предусмотрено условие остановки по достижении определенного числа k переменных в выходном наборе. Тогда, например, для процедур прямого присоединения и присоединения-удаления достаточно положить $k = p$. В случае условия остановки, управляемого величиной $F_{вкл}$, увеличения числа отбираемых переменных можно добиться, уменьшая

значение $F_{\text{вкл}}$, полагая его равным 20% или даже 30% уровню значимости.

На рис. 8.1 приведены два графика критерия качества набора для процедуры последовательного присоединения (значения критерия качества определены лишь в целых точках, однако для наглядности они соединены линией). Кривая I отражает случай, наиболее часто возникающий при отборе переменных: сначала монотонное возрастание величины критерия качества, а затем ее монотонное убывание. Набор, со-

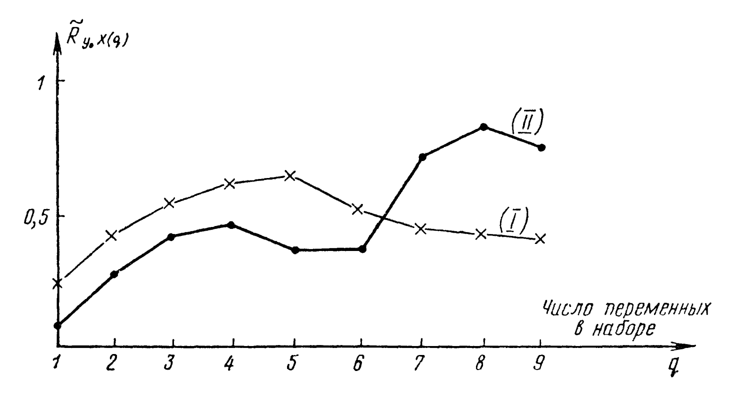


Рис. 8.1. Варианты зависимости несмещенной оценки коэффициента множественной корреляции ($\widetilde{R}_{y.X(q)}$) от количества переменных для пошаговой процедуры последовательного присоединения

ответствующий точке максимума, или какой-либо набор в ближайшей (плюс—минус одна-две переменные) его окрестности, является искомым информативным набором. Кривая II представляет потенциально более интересный случай отбора: после достижения локального минимума кривая вновь начинает возрастать, и величина критерия качества даже превосходит первый максимум. В этом случае целесообразно исследовать следующие вопросы:

добавление какой переменной изменило ход графика?

пусть это переменная $x^{(j_l)}$, тогда сочетание каких переменных из $X_{-j_l}(l-1)$ и $x^{(j_l)}$ привело к скачку критерия качества? В первую очередь подозрительна переменная $x^{(j_{l-1})}$. Затем необходимо провести отбор переменных с принудительным включением переменных $x^{(j_l)}$, $x^{(j_{l-1})}$ и других переменных, обусловивших изменение хода графика (такой отбор может также использоваться и для получения ответа на второй во-

прос, если, кроме $x^{(i_{l-1})}$, в изменении хода графика «виновны» еще и другие переменные из $X_{-i_l}(l-1)$).

2. Экспертное упорядочение переменных по степени их информативности. Для успешного применения процедур отбора, в особенности когда переменных много, важную роль играет априорная (экспертная) оценка значимости потенциальных переменных для рассматриваемой задачи [2, 3, 93]. Например, источником для такой априорной оценки могут быть, во-первых, содержательные соображения об исследуемом явлении и, во-вторых, задачи-аналоги, с которыми уже имел дело исследователь. Во всяком случае полезно разделить имеющиеся переменные на три группы ([93, гл. 15]): 1) *ключевые*—переменные, о которых известно, что они оказывают существенное влияние на зависимую переменную y; все или некоторые из этих переменных могут быть по требованию исследователя включены в выходной набор в принудительном порядке; 2) *потенциально информативные* — переменные, возможность влияния которых на зависимую переменную y представляется достаточно обоснованной; 3) «*шумовые*» — переменные, влияние которых на переменную представляется маловероятным.

После сортировки переменных отбор производится следующим образом.

На первом этапе задача регрессии решается в пространстве ключевых переменных. Проводится анализ точности и адекватности соответствующей линейной модели (см. гл. 11). Если не все из ключевых переменных необходимо в принудительном порядке включить в итоговую модель, то можно попытаться сократить их число, применяя тот или иной пошаговый алгоритм. При этом переменные, не вошедшие в информативный набор, переводятся в группу потенциально информативных переменных.

Второй этап проводится, если качество регрессионного уравнения, оцененного на первом этапе, является неудовлетворительным. На этом этапе осуществляется отбор переменных из множества, полученного объединением ключевых и потенциально информативных переменных. Переменные, отобранные на первом этапе, включаются в выходной набор в обязательном порядке. Переменные, не вошедшие в информативный набор на втором этапе, переводятся в группу «шумовых». Если первые два этапа не привели к удовлетворительному результату, проводится отбор среди «шумовых» переменных с принудительным включением переменных, отобранных на первом и втором этапах.

ВЫВОДЫ

1. При практическом применении мнк-оценок исследователь часто сталкивается с явлением мультиколлинеарности, когда объясняющие переменные сильно коррелированы, т. е. существуют выраженные, хотя и неточные, линейные связи между несколькими или всеми объясняющими переменными. В этой ситуации точность обычных мнк-оценок резко падает: ошибки некоторых параметров уравнения регрессии становятся очень большими, эти ошибки сильно скоррелированы, выборочные дисперсии резко возрастают. Резко сокращаются возможности интерпретации уравнения регрессии. Степень мультиколлинеарности измеряется либо обратной величиной минимального собственного числа нормированной (корреляционной) матрицы, либо числом обусловленности, равным отношению максимального собственного числа к минимальному. Если минимальное собственное число равно нулю, то степень мультиколлинеарности и число обусловленности являются бесконечно большими, и мы имеем дело с точной мультиколлинеарностью или вырожденной системой линейных уравнений.

2. Оценивание параметров уравнения регрессии в случае сильной мультиколлинеарности основано на различных методах регуляризации задачи — модификациях регрессии на главные компоненты, гребневых и редуцированных оценках. Со статистической точки зрения получаемые оценки являются, в отличие от мнк-оценок, смещенными. Однако они обладают рядом оптимальных свойств, в частности обеспечивают лучшие прогностические свойства оцененного уравнения регрессии на объектах, не вошедших в обучающую выборку.

3. Одним из методов получения оценок параметров уравнения регрессии при мультиколлинеарности является отбор существенных (информативных) объясняющих переменных. Существует ряд мер качества набора переменных, которые используются алгоритмами отбора. Все они являются функциями от коэффициента детерминации, объема выборки и количества переменных, входящих в набор. В отличие от коэффициента детерминации, который не может уменьшаться при расширении набора объясняющих переменных, меры качества, используемые при отборе переменных, могут при этом убывать.

4. Алгоритмы отбора переменных отличаются используемым критерием качества набора и способом генерации наборов переменных для их сравнения. Из схем генерации удобными с вычислительной точки зрения являются пошаговые схемы — простого добавления, простого удаления, добавления с удалением и схемы выметания. В настоящее время в связи с рос-

том возможностей ЭВМ получают распространение и схемы прямого перебора, и различные его оптимизации на основе метода ветвей и границ. Пошаговые схемы хотя и не гарантируют получения оптимального по выбранному критерию набора, однако позволяют обычно получить наборы, вполне удовлетворительные для практического применения.

Глава 9. ВЫЧИСЛИТЕЛЬНЫЕ АСПЕКТЫ МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

9.1. Итерационные методы поиска оценок метода наименьших квадратов (мнк-оценок)

9.1.1. Введение. Гл. 7, 8, 10 и 11 в той или иной мере посвящены изучению статистических свойств оценок, порождаемых регрессионными моделями. В них рассмотрены такие вопросы, как состоятельность, несмещенность, эффективность и т. п. В настоящей главе все внимание уделено численным процедурам отыскания оценок. Для отыскания большинства оценок (см. гл. 5, 7) приходится решать экстремальные задачи вида

$$\widehat{\Theta}^* = \arg\min_{\Theta \in \Gamma} \sum_{i=1}^{n} [Y_i - \mathbf{f}(X_i; \Theta)]' \, \mathbf{W}_i \, [Y_i - \mathbf{f}(X_i; \Theta)]. \tag{9.1}$$

Обычно на подобные задачи ссылаются как на задачи *взвешенного метода наименьших квадратов*. В дальнейшем (9.1) будет именоваться задачей мнк и рассматриваться лишь случай *одномерного* отклика.

Итак, ближайшая цель — изложение численных методов решения экстремальной задачи

$$\widehat{\Theta}^* = \arg\min_{\Theta \in \Gamma} \sum_{i=1}^{n} w_i (y_i - \mathbf{f}(X_i; \Theta))^2. \tag{9.2}$$

Для некоторого упрощения записей там, где это не вызовет разночтений, вместо $f(X_i; \Theta)$ будет использоваться запись $f_i(\Theta)$.

Можно рассматривать (9.2) как задачу нелинейного программирования:

$$\widehat{\Theta}^* = \arg\min_{\Theta \in \Gamma} J(\Theta), \tag{9.3}$$

где $J(\Theta) = \sum_{i=1}^{n} w_i (y_i - f_i(\Theta))^2$. Многочисленные стандартные

алгоритмы и программы для решения задач нелинейного программирования имеются в математическом обеспечении практически любой современной ЭВМ (см. гл. 15).

Особенности использования общих алгоритмов минимизации в статистических задачах неоднократно обсуждались в литературе (см., например, [172], эта работа содержит обширную библиографию; [145, 146, 135, 43]). Вообще говоря, любой из этих алгоритмов пригоден для решения (9.3), однако имеются веские аргументы для развития специальных алгоритмов и программ решения экстремальных задач, связанных с анализом данных регрессионных экспериментов.

Во-первых, *учет структуры* функции $J(\Theta)$ позволяет отобрать алгоритмы, работающие наиболее эффективно именно при решении задач типа (9.2).

Во-вторых, решение экстремальной задачи (9.2) или (9.3) составляет примерно лишь половину от общего объема вычислительной работы, которую необходимо проделать при регрессионном анализе. Действительно, сами оценки, т. е. числа $\widehat{\Theta}^*$, содержат не так уж много информации для исследователя. *Необходимо знать ковариационные матрицы оценок или их оценки*, доверительные интервалы для неизвестного значения результирующего показателя (отклика), ряд величин, характеризующих адекватность регрессионной модели и т. д. (см. гл. 11). В вычислительном плане крайне удобно, когда упомянутая числовая информация подсчитывается как «побочная» при отыскании самих оценок. Именно алгоритмам, обладающим такими свойствами, отдается предпочтение при создании программ по регрессионному анализу.

9.1.2. Алгоритмы квазиградиентного типа. Предположим, что область допустимых значений параметров Γ совпадает со всем евклидовым пространством R^m (m — число неизвестных параметров, т. е. размерность вектора Θ).

Наибольшее распространение в настоящее время получили алгоритмы итерационного типа

$$\widehat{\Theta}_{s+1} = \widehat{\Theta}_s + \rho_s \delta_s, \tag{9.4}$$

где s — номер итерации; δ_s — вектор, определяющий направление движения на s-й итерации; ρ_s — длина шага.

Идея, лежащая в основе этих алгоритмов, очень проста: на каждом шаге двигаться в направлении минимума функции $J(\Theta)$. Различные алгоритмы отличаются способом выбора этого направления и правилами выбора длины шага. В данной главе обсуждаются лишь алгоритмы, движение в которых осуществляется в направлении под острым углом к антиградиен-

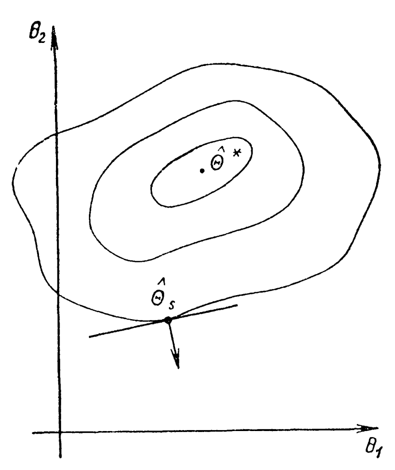


Рис. 9.1. Определение направления градиента

ту—$\nabla J(\Theta)$ функции $J(\Theta)$ (или некоторой ее аппроксимации). Такие алгоритмы будут называться в дальнейшем алгоритмами *квазиградиентного типа*. Напомним, что антиградиент—это направление, противоположное градиенту, а градиент в точке Θ перпендикулярен к линии постоянного значения функции $J\Theta$), проходящей через эту точку (рис. 9.1). Полезно иметь в виду, что градиент определяется часто следующим образом:

$$\nabla J(\Theta) = \max_{\mathbf{e}} \lim_{\lambda \to 0} \frac{J(\Theta + \lambda \mathbf{e})}{\lambda},$$

где $\mathbf{e}'\mathbf{e} = 1$.

Таким образом, градиент в каком-то смысле указывает направление локального наискорейшего возрастания функции $J(\Theta)$. Компоненты градиента определяются формулой

$$\nabla J(\Theta) = \frac{\partial}{\partial \Theta} J(\Theta) = \left(\frac{\partial}{\partial \Theta_1} J(\Theta), \ldots, \frac{\partial}{\partial \Theta_m} J(\Theta) \right).$$

Легко проверить, что для функции $J(\Theta)$, соответствующей задаче мнк,

$$\Delta J(\Theta) = -2\mathcal{Y}(\Theta),$$

где

$$\mathcal{Y}(\Theta) = \sum_{i=1}^{n} w_i (y_i - f_i(\Theta)) \frac{\partial f_i(\Theta)}{\partial \Theta}.$$

По рисунку видно, что в некоторых точках направление, указываемое антиградиентом, существенно отличается от направления, указывающего на точку $\widehat{\Theta}^*$. По этой причине многие алгоритмы предусматривают движение, вообще говоря, отличное от антиградиента.

Для алгоритмов квазиградиентного типа соотношение (9.4) принимает вид:

$$\hat{\Theta}_{s+1} = \hat{\Theta}_s - \rho_s \mathrm{H}_s \tilde{\nabla} J_s \tag{9.5}$$

Вектор $\tilde{\nabla} J_s$ либо совпадает с градиентом $\nabla J(\hat{\Theta}_s)$, подсчитанным в точке $\hat{\Theta}_s$, либо представляет из себя его некоторую аппроксимацию. Матрица $\mathbf{H}_s$ — положительно полуопределенная матрица, т. е. $\tilde{\nabla} J_s' \mathbf{H}_s \tilde{\nabla} J_s \geqslant 0$, что и гарантирует движение под острым углом к антиградиенту.

В табл. 9.1 представлены выражения для наиболее распространенных алгоритмов безусловной минимизации

Таблица 9.1

№ п/п	Метод	Матрица $\mathbf{H}_s$	Пункт книги, где описан способ выбора	Тип скорости сходимости
1	Градиентный спуск	$\mathbf{H}_s \equiv \mathbf{I}_m$	9.2.1	Геометрическая прогрессия
2	Ньютона	$\frac{1}{2}[\mathbf{M}_s + \mathbf{V}_s]^{-1}$	9.3.1	Сверхлинейная (квадратичная)
3	Ньютона — Гаусса (линеаризации)	$\frac{1}{2}\mathbf{M}_s^{-1}$	9.4.1	Геометрическая прогрессия
4	Вариант Марквардта	$\frac{1}{2}[\mathbf{M}_s + \gamma_s \mathbf{A}_s]^{-1},\ \mathbf{A}_s > 0$	9.4.2	Геометрическая прогрессия
5	Сопряженных градиентов	$\mathbf{I}_m - \frac{g_s \nabla J'(\Theta_s)\mathbf{H}_s}{\nabla J'(\Theta_s)\nabla J(\Theta_s)},\ \mathbf{H}_0 = \mathbf{I}_m$	9.5.3	Сверхлинейная

9.2. Градиентный спуск

9.2.1. Описание общей схемы алгоритма. При градиентном спуске движение осуществляется непосредственно в направлении антиградиента, т. е. $\mathbf{H}_s = \mathbf{I}_m$ (напомним, что $\mathbf{I}_m$ — единичная матрица размерности $m \times m$). Итерационная процедура таким образом принимает вид:

$$\hat{\Theta}_{s+1} = \hat{\Theta}_s + \frac{1}{2}\rho_s \mathcal{Y}_s, \tag{9.6}$$

где $\mathcal{Y}_s = \mathcal{Y}(\widehat{\Theta}_s)$.

Перечислим несколько возможных способов выбора величины шага ρ_s.

Обозначим $p_s = \frac{1}{2}\mathcal{Y}_s$ направление минимизации. Существуют два основных способа, приводящих к снижению значения $J(\Theta)$ на каждом шаге и к сходимости итеративного процесса.

1. Зададимся некоторым $0 < \varepsilon < 1$. Дроблением шага добьемся того, чтобы

$$J(\widehat{\Theta}_{s+1}) - J(\widehat{\Theta}_s) \leqslant \varepsilon p_s' \nabla J(\widehat{\Theta}_s).$$

Поскольку $p_s' \nabla J(\widehat{\Theta}_s) < 0$, всегда $J(\widehat{\Theta}_{s+1}) - J(\widehat{\Theta}_s) < 0$.

2. Длина шага определяется из условия

$$\rho_s = \arg\min_{\rho \geqslant 0} J(\widehat{\Theta}_s + \rho\, p_s). \qquad (9.7)$$

При таком выборе шага обычно говорят о «наискорейшем спуске». Экстремальная задача (9.7) чаще всего решается с помощью квадратичной аппроксимации по ρ.

Решение одномерной задачи минимизации вторым способом может оказаться чрезвычайно трудоемким. Дело может осложниться тем, что функция $J(\Theta)$ вдоль выбранного направления может быть *мультимодальной*. Поэтому первый способ нам кажется более предпочтительным. Описанные способы выбора шага могут применяться и в других методах минимизации (см. § 9.3—9.5).

Сравнение эффективности различных способов выбора длины шага применительно к задачам регрессии проведено в [159].

Алгоритмы типа (9.6) (см., например, [35, 107, 25]) обеспечивают при определенных ограничениях на функцию $J(\Theta)$ сходимость последовательности $\{\widehat{\Theta}_s\}$ со скоростью геометрической прогрессии

$$\|\widehat{\Theta}_s - \widehat{\Theta}^*\| \leqslant Cq^s \qquad (9.8)$$

В частности, такая скорость сходимости обеспечивается так называемой *линейной сходимостью*, при которой

$$\|\widehat{\Theta}_{s+1} - \widehat{\Theta}^*\| \leqslant q \|\Theta_s - \Theta^*\|, \quad 0 < q < 1,$$

где $\|\widehat{\Theta}_s - \widehat{\Theta}^*\|$ — длина вектора $\Theta_s - \Theta^*$; C и q — константы, определяемые видом $J(\Theta)$.

Например, если помимо некоторых не очень существенных ограничений градиент удовлетворяет условию Липшица:

$$\| \nabla J(\Theta) - \nabla J(\widetilde{\Theta}) \| \leqslant L \| \Theta - \widetilde{\Theta} \|$$

при всех Θ, $\widetilde{\Theta} \in R^m$, $L = \text{const} > 0$, и функция $J(\Theta)$ сильно выпукла с показателем μ:

$$J[\alpha\Theta + (1-\alpha)\widetilde{\Theta}] \leqslant \alpha J(\Theta) + (1-\alpha) J(\widetilde{\Theta}) -$$

$$-\mu\alpha(1-\alpha)(\Theta - \widetilde{\Theta})'(\Theta - \widetilde{\Theta})$$

при всех Θ, $\widetilde{\Theta} \in R^m$ и $0 \leqslant \alpha \leqslant 1$, то $q = 1 - \mu/2L$.

9.2.2. Замечание об эффективности алгоритма. Одним из основных достоинств градиентного спуска является его простота. Однако реальная скорость его сходимости уменьшается при приближении Θ_s к точке Θ^*. Для функций овражного типа с сильно вытянутыми линиями уровня в окрестности Θ^* эффективность методов типа градиентного спуска особенно низка, так как обычно для таких функций μ близко к нулю.

При решении статистических задач с помощью градиентного спуска приходится на заключительном этапе проводить дополнительные расчеты по отысканию оценок ковариационных матриц и прочих величин, описывающих статистические свойства оценок.

Обычно градиентный спуск целесообразно применять лишь на начальных этапах минимизации, используя найденные в результате сравнительно небольшого числа итераций величины $\widehat{\Theta}_s$ в качестве начального приближения для более сложных методов, обладающих большей скоростью сходимости.

9.3. Метод Ньютона

9.3.1. Описание общей схемы метода. Идея метода Ньютона (иногда его называют методом Ньютона—Рафсона) заключается в квадратичной аппроксимации функции $J(\Theta)$ в окрестности точки $\widehat{\Theta}_s$. Значения $\widehat{\Theta}_{s+1}$ находятся из условия минимума аппроксимирующего полинома второй степени и определяются в случае положительной определенности матрицы

$$\mathbf{G}_s = \left. \frac{\partial^2 J(\Theta)}{\partial \Theta \, \partial \Theta'} \right|_{\Theta = \hat{\Theta}_s},$$

по формуле

$$\widehat{\Theta}_{s+1} = \widehat{\Theta}_s - \mathbf{G}_s^{-1} \nabla J(\widehat{\Theta}_s). \tag{9.9}$$

Положительная определенность *гессиана* $\mathbf{G}_s$ является существенным ограничением использования метода Ньютона. Вместе с тем, чем ближе начальное приближение к минимуму, тем скорее можно ожидать выполнение этого условия. Ведь в точке минимума, весьма вероятно, матрица $\mathbf{G}(\widehat{\Theta}^*)$ положительно определена, а из непрерывности $\mathbf{G}(\Theta)$ следует, что в некоторой окрестности $\widehat{\Theta}^*$ гессиан также будет положительно определен. Поэтому наибольший эффект имеет применение этого метода в достаточно близкой окрестности решения.

Иными словами, $\rho_s = 1$, $\mathbf{H}_s = \mathbf{G}_s^{-1}$. Несложные выкладки показывают, что для (9.2)

$$\mathbf{G}_s = 1/2\,(\mathbf{M}_s + \mathbf{V}_s),$$

где

$$\mathbf{M}_s = \sum_{i=1}^{n} w_i \dot{\mathbf{f}}_{is} \dot{\mathbf{f}}_{is}'; \qquad \dot{\mathbf{f}}_{is} = \left.\frac{\partial f_i(\Theta)}{\partial\Theta\,\partial\Theta'}\right|_{\Theta=\Theta_s};$$

$$\mathbf{V}_s = -\sum_{i=1}^{n} w_i (y_i - f_i(\Theta_s))\,\mathbf{\Phi}_{is}; \qquad \mathbf{\Phi}_{is} = \left.\frac{\partial^2 f_i(\Theta)}{\partial\Theta\,\partial\Theta'}\right|_{\Theta=\Theta_s}$$

При линейной параметризации $f_i(\Theta) = \Theta'\mathbf{f}^0(X_i)$ решение $\widehat{\Theta}^*$ получается на первом же шаге независимо от выбора Θ_0. На практике предпочитают использовать *метод Ньютона с регулировкой шага*

$$\Theta_{s+1} = \Theta_s + \rho_s (\mathbf{M}_s + \mathbf{V}_s)^{-1} \mathcal{Y}_s, \tag{9.10}$$

где ρ_s выбирается, например, в соответствии со способом 1 из предыдущего параграфа или из условия

$$\rho_s = \arg\min_{\rho>0} J[\Theta_s + \rho\,(\mathbf{M}_s + \mathbf{V}_s)^{-1} \mathcal{Y}_s].$$

Процедура (9.10) оказывается более стабильной по сравнению с (9.9), которая особенно чувствительна к выбору начального приближения Θ_0 и подвержена эффекту «раскачки» при его неудачном выборе.

9.3.2. Скорость сходимости процедуры. Если дополнительно к условиям, сформулированным в конце п. 9.2.1, потребовать, чтобы $\|\mathbf{G}(\Theta) - \mathbf{G}(\widehat{\Theta})\| \leqslant K\|\Theta - \widehat{\Theta}\|$ при всех Θ, $\widehat{\Theta} \in R^m$, то при упомянутых последовательностях $\{\rho_s\}$ независимо от выбора Θ_0 последовательность $\{\Theta_s\}$ сходится к $\widehat{\Theta}^*$ с *квадратичной скоростью*, т. е.

$$\|\widehat{\Theta}_{s+1} - \widehat{\Theta}^*\| \leqslant C\|\widehat{\Theta}_s - \widehat{\Theta}^*\|^2,$$

где константа C определяется видом функции $J(\Theta)$ и не зависит от s.

При решении практических задач на данное утверждение (впрочем, как и на аналогичное утверждение из § 9.2) не следует особенно полагаться. Дело в том, что проверка условий, его сопровождающих, за исключением тривиальных случаев, реально невозможна. К тому же большинство из них для «экзотических» выборок (маловероятных выборок) заведомо не будут выполняться. Тем не менее подобные утверждения все же имеют смысл, так как позволяют дать оценку той максимальной скорости сходимости, которую можно достигнуть с помощью данного метода. Данное замечание имеет место для всех методов, рассматриваемых в этой главе.

Одним из наиболее существенных недостатков метода Ньютона является необходимость подсчета производных $\dot{\mathbf{I}}_{is}$ и $\mathbf{\Phi}_{is}$. Для достаточно сложных функций $f_i(\Theta)$ это приводит к весьма громоздким вычислениям и заметно усложняет работу пользователя, так как приходится составлять специальные дополнительные программы по подсчету производных.

9.4 Метод Ньютона-Гаусса и его модификации

9.4.1. Общая схема метода. Заметно более простым по сравнению с предыдущим методом является метод Ньютона—Гаусса, в котором матрица $\mathbf{H}_s = \mathbf{M}_s^{-1}$. Практика показывает, что именно для регрессионных задач его эффективность такая же, как и метода Ньютона.

К итерационной процедуре Ньютона—Гаусса

$$\widehat{\Theta}_{s+1} = \widehat{\Theta}_s + \rho_s \mathbf{M}_s^{-1} \mathcal{Y}_s \qquad (9.11)$$

можно прийти из следующих соображений. Для достаточно гладких функций $f_i(\Theta)$ в окрестности точки $\widehat{\Theta}_s$ можно полагаться на простейшую аппроксимацию

$$f_i(\Theta) = f_i(\widehat{\Theta}_s) + \dot{\mathbf{f}}'_{is}(\Theta - \widehat{\Theta}_s). \qquad (9.12)$$

Полагая $\widetilde{y}_i = y_i - f_i(\widehat{\Theta}_s)$ и $\widetilde{\Theta} = \Theta - \widehat{\Theta}_s$, приходим к необходимости минимизации (см. (9.2)) функции

$$\sum_{i=1}^{n} w_i (\widetilde{y}_i - \dot{\mathbf{f}}'_{is} \widetilde{\Theta})^2.$$

При исследовании этой задачи в гл. 7 показано, что минимум достигается при $\widetilde{\Theta} = \mathbf{M}_s^{-1}\,\mathcal{Y}_s$. Отсюда следует, что $\widehat{\Theta}^* \approx \widehat{\Theta}_s + \mathbf{M}_s^{-1}\,\mathcal{Y}_s$.

Для линейного случая решение достигается за один шаг. При нелинейной параметризации процедура повторяется:

$$\widehat{\Theta}_{s+1} = \widehat{\Theta}_s + \mathbf{M}_s^{-1}\,\mathcal{Y}_s. \tag{9.13}$$

Именно эта процедура и носит название метода Ньютона—Гаусса.

Для рассматриваемой экстремальной задачи метод Ньютона—Гаусса близок методу Ньютона. При линейной параметризации они совпадают. Их близость при малых вторых производных $\mathbf{\Phi}_{is}$ очевидна. Имеется и более глубокая причина их близости. Действительно, при $n \to \infty$ и некоторых не слишком ограничительных предположениях в силу закона больших чисел имеем следующую сходимость (с вероятностью единица)

$$n^{-1}\sum_{i=1}^{n} w_i\,(y_i - f_i\,(\widehat{\Theta}_s))\,\mathbf{\Phi}_{is} \to \lim_{n} n^{-1}\sum_{i=1}^{n} w_i\,(f_i\,(\Theta_\text{и}) - f_i\,(\widehat{\Theta}_s))\,\mathbf{\Phi}_{is},$$

где $\Theta_\text{и}$ — истинные значения искомых параметров.

9.4.2. Обсуждение скорости сходимости процедуры. Метод Ньютона—Гаусса очень чувствителен к обусловленности матриц $\mathbf{M}_s$. При плохо обусловленных матрицах $\mathbf{M}_s$ наблюдается «раскачка» итерационного процесса, а если он и сходится, то его предельные точки меняются с изменением начального приближения Θ_0. Наиболее распространенной причиной плохой обусловленности матриц $\mathbf{M}_s$ является неудачный выбор режимов наблюдений $\mathbf{X}$. Поэтому, сталкиваясь с плохо обусловленными матрицами $\mathbf{M}_s$, экспериментатору следует попытаться в первую очередь разобраться в своих опытных данных, и, может быть, провести дополнительные наблюдения. Если же структура данных не может быть улучшена, то приходится обращаться к методам, которые менее чувствительны к виду матриц $\mathbf{M}_s$. Одним из наиболее широко применяемых является метод Марквардта:

$$\widehat{\Theta}_{s+1} = \widehat{\Theta}_s + \rho_s\,[\mathbf{M}_s + \gamma_s\,\mathbf{A}_s]^{-1}\,\mathcal{Y}_s, \tag{9.14}$$

который может трактоваться как некоторое усовершенствование метода Ньютона—Гаусса.

В (9.14) $\gamma_s \geqslant 0$, $\mathbf{A}_s$ — положительно полуопределенная матрица. При $\gamma_s = 0$ реализуется метод Ньютона—Гаусса, при $\gamma_s \to \infty$ и $\mathbf{A}_s = \mathbf{I}$ направление движения приближается к антиградиенту. Выбор ρ_s и γ_s в большинстве модификаций (9.14) проводится из соображений монотонного убывания $J(\Theta)$.

Матрица $\mathbf{A}_s$ в большинстве компьютерных реализаций (9.14) выбирается диагональной, причем ее элементы совпадают с диагональными элементами матрицы $\mathbf{M}_s$.

Полезно иметь в виду следующий факт. Если опираться на линейную аппроксимацию (9.12), то при $\rho_s = 1$ каждый шаг в методе Марквардта может быть истолкован как минимизация функции

$$\sum_{i=1}^{n} w_i (\tilde{y}_i - \mathbf{f}'_{is}\tilde{\Theta})^2 + \gamma_s \tilde{\Theta}' \mathbf{A}_s \tilde{\Theta}.$$

Иными словами, в этом методе на каждом шаге проводится регуляризация исходной задачи.

Сходимость метода Ньютона—Гаусса и его модификаций изучалась, например, в [109, 200, 237], различные комментарии и дополнительную библиографию можно найти в [145, 146, 25, 43]. Скорость сходимости в зависимости от условий, накладываемых на функции $f_i(\Theta)$, $\mathbf{f}_{is}$, $\mathbf{\Phi}_{is}$, и способов выбора ρ_s, γ_s, $\mathbf{A}_s$ может быть линейной ($\|\widehat{\Theta}_{s+1} - \widehat{\Theta}^*\| \leqslant q\|\widehat{\Theta}_s - \widehat{\Theta}^*\|$), сверхлинейной ($\|\widehat{\Theta}_{s+1} - \widehat{\Theta}^*\| \leqslant q_s\|\widehat{\Theta}_s - \widehat{\Theta}^*\|$, $q_s \to 0$) или квадратичной ($\|\widehat{\Theta}_{s+1} - \widehat{\Theta}^*\| \leqslant C\|\widehat{\Theta}_s - \widehat{\Theta}^*\|^2$).

9.4.3. Рекомендации по правилу остановки итерационной процедуры. Для регрессионной задачи (9.2) матрица $\mathbf{D} = \mathbf{M}^{-1}(\widehat{\Theta}^*)$ может быть использована в качестве оценки ковариационной матрицы мнк-оценок. Если $\lim_{s\to\infty} \widehat{\Theta}_s = \widehat{\Theta}^*$, то итерационная процедура Ньютона—Гаусса наряду с оценками $\widehat{\Theta}^*$ поставляет и матрицу $\mathbf{D}$ ($\mathbf{D} \approx \mathbf{M}_{s^*}^{-1}$, где s^* — номер заключительной итерации). Этот факт позволяет также сформулировать простое и естественное правило остановки. Расчеты прекращаются, как только

$$(\widehat{\Theta}_{s+1} - \widehat{\Theta}_s)' \mathbf{M}_{s+1}^{-1} (\widehat{\Theta}_{s+1} - \widehat{\Theta}_s) \leqslant \alpha, \tag{9.15}$$

где α — наперед заданная точность.

Данное правило более естественно, чем, например, правило, используемое в общих программах минимизации: $(\widehat{\Theta}_{s+1} - \widehat{\Theta}_s)' (\widehat{\Theta}_{s+1} - \widehat{\Theta}_s) \leqslant \alpha$. Действительно, точность нахождения Θ^* целесообразно соизмерять со статистической точностью (ее ковариационной матрицей) мнк-оценок.

9.5. Методы, не использующие вычисления производных

9.5.1. Основные подходы к устранению необходимости вычисления производных. Как уже отмечалось, существенным недостатком методов, изложенных в предыдущих параграфах, является необходимость подсчета производных $\mathbf{f}_{is}$, а в методе Ньютона — и вторых производных $\mathbf{\Phi}_{is}$ на каждой итерации. При сложных функциях $f_i(\Theta)$ это, во-первых, оказывается утомительным с программистской точки зрения, а во-вторых, приводит к громоздким вычислениям на каждой итерации.

Возможно несколько способов избавления от необходимости подсчета производных:

использование методов прямого поиска (нулевого порядка), таких, как симплекс-метод, метод случайного поиска и т. п. [185],

аппроксимация производных конечно-разностными аналогами,

специальные методы аппроксимации матриц $\mathbf{H}_s$, ∇J_s, позволяющие реализовать итерационные процедуры, близкие по эффективности к процедурам Ньютона и Ньютона—Гаусса, но с меньшим объемом вычислений.

Прямые методы минимизации оказались малоэффективными для задач регрессионного типа даже по сравнению с градиентными методами [25]. К тому же после отыскания с их помощью экстремальных значений $\widehat{\Theta}^*$ требуется проведение объемистых вычислений по нахождению статистических характеристик, описывающих качество оценок.

Конечно-разностная аппроксимация хотя и избавляет потребителя от утомительной работы по составлению дополнительных программ для вычисления производных, но не решает проблемы сокращения объема вычислений. Чаще всего она приводит к его заметному увеличению.

Наиболее перспективными и удобными оказались методы третьей группы. Они завоевали весьма прочное место во многих статистических пакетах (см., например, [169]).

9.5.2. Разностные аналоги метода Ньютона — Гаусса. Основная идея, на которую опираются методы третьей группы, заключается в использовании на $(s+1)$-й итерации информации, полученной на предыдущих s итерациях, для построения разумных аппроксимаций элементов матрицы $\mathbf{H}_s$ и компонент градиента ∇J_s.

При решении регрессионных задач хорошо зарекомендовали себя методы, являющиеся, по существу, аппроксимациями метода Ньютона—Гаусса. По-видимому, работа [236] является в этом направлении пионерской. Упрощенный и непосредственно приспособленный к задачам регрессии алгоритм предложен в [242], см. также [169]. Достаточно подробный анализ алгоритмов подобного типа и их дальнейшее развитие содержатся в [37, 38]. Данную совокупность методов целесообразно назвать методами Ньютона—Гаусса без подсчета производных. Как и обычный метод Ньютона—Гаусса (см. п. 9.4.2), эти методы опираются на линейную аппроксимацию функций $f_i(\Theta)$ в окрестности точки $\widehat{\Theta}_s$:

$$f_i(\Theta) \approx f_i(\widehat{\Theta}_s) + \gamma'_{is}(\Theta - \widehat{\Theta}_s). \tag{9.16}$$

В отличие от метода Ньютона—Гаусса коэффициенты γ_{is} не совпадают с производными $\dot{\mathbf{f}}'_{is}$, а подсчитываются по значениям функции $f_i(\Theta_t)$, полученным по прошлым итерациям. В принципе величины γ_{is} могут быть подсчитаны самыми различными способами, но в программах по регрессионному анализу оказывается очень удобным отыскивать их с помощью все того же мнк. Определим γ_{is} как решение задачи мнк:

$$\gamma_{is} = \arg\min_{\gamma} \sum_{t=1}^{s+q} \omega_{st} [f_i(\widehat{\Theta}_t) - f_i(\widehat{\Theta}_s) - \gamma'(\widehat{\Theta}_t - \widehat{\Theta}_s)]^2,$$

где ω_{st} — веса, описывающие вклад того или иного значения $f_i(\widehat{\Theta}_t)$. При этом s значений получены непосредственно из итерационной процедуры, а выбор остальных q значений парамет ров Θ будет объяснен позднее. В частности, в алгоритме, предложенном в [242], $\omega_{s,s+q} = \omega_{s,s+q-1} = \ldots = \omega_{s,s+q-m} = 1$, а остальные веса полагаются нулевыми.

Хорошие результаты дает выбор весов вида $\omega_{st} = \varphi[J(\Theta_t)]$, где φ — убывающая функция, например $\exp[-kJ(\Theta_t)]$.

Можно показать, что

$$\gamma_{is} = \mathbf{Q}_s^{-1} U_{is}, \tag{9.17}$$

где

$$\mathbf{Q}_s = \sum_{t=1}^{s+q} \omega_{st} \Delta_{st} \Delta'_{st}; \quad U_{is} = \sum_{t=1}^{s+q} \omega_{st} \Delta_{st} u_{ist};$$

$$\Delta_{st} = \widehat{\Theta}_t - \widehat{\Theta}_s; \quad u_{ist} = f_i(\widehat{\Theta}_t) - f_i(\widehat{\Theta}_s).$$

Подставляя теперь в (9.2) приближенное значение функции $f_i(\Theta) = f_i(\widehat{\Theta}_s) + \gamma'_{is}(\Theta - \widehat{\Theta}_s)$ (ср. с § 9.4), получим, что функция $J(\Theta)$ достигает своего минимума при $\Theta - \widehat{\Theta}_s = \mathbf{M}_s^{-1} \mathcal{Y}_s$,

где

$$\mathbf{M}_s = \sum_{i=1}^{n} w_i \gamma_{is} \gamma'_{is}; \quad \mathcal{Y}_s = \sum_{i=1}^{n} w_i \gamma_{is} (y_i - f_i(\widehat{\Theta}_s)).$$

После несложных преобразований приходим к очень удобной в вычислительном плане формуле

$$\Theta - \widehat{\Theta}_s = \mathbf{Q}_s \mathbf{m}_s^{-1} \sum_{i=1}^{n} U_{is} (y_i - f_i(\widehat{\Theta}_s)), \tag{9.18}$$

где $\mathbf{m}_s = \sum_{i=1}^{n} w_i U_{is} U'_{is}$

Итерационная процедура Ньютона—Гаусса, опирающаяся на (9.18), принимает вид

$$\widehat{\Theta}_{s+1} = \widehat{\Theta}_s + \rho_s \mathbf{Q}_s \mathbf{m}_s^{-1} \sum_{i=1}^{n} U_{is} (y_i - f_i(\Theta_s)). \tag{9.19}$$

Выбор ρ_s осуществляется по одному из правил, принимаемых для процедуры Ньютона—Гаусса с переменным шагом.

Остановимся на некоторых особенностях процедуры (9.19). Ее основное достоинство состоит в том, что *на каждом шаге достаточно всего одного вычисления функции* $f_i(\Theta)$ (естественно, при всех $i = \overline{1, n}$). Данное свойство оказывается чрезвычайно полезным при сложных функциях $f(X_i; \Theta)$, например, в тех случаях, когда эта функция удовлетворяет некоторому дифференциальному уравнению, допускающему лишь численное решение, для получения которого требуются специальные объемистые вычисления.

Процедура (9.19) содержит операцию обращения матрицы, от которой можно было бы избавиться, заменив ее обращением

с помощью рекуррентных формул. Однако, как правило, трудоемкость этой операции несущественна по сравнению с трудоемкостью подсчета значений функций f_i (Θ).

В [242] утверждается, что вместо непосредственного обращения матрицы $\mathbf{m}_s$ целесообразно использовать пошаговые процедуры, применяемые при отборе существенных факторов (см. гл. 8). Подобный прием особенно удобен при функциях J (Θ) «овражного» типа (матрица $\mathbf{m}_s$ плохо обусловлена).

Для определения $\boldsymbol{\gamma}_{i1}$ требуется подсчет функций f_i (Θ) по крайней мере в (m + 1)-й точке Θ_l. На последующих итерациях достаточно проведения подсчета функций f_i (Θ) лишь в одной точке Θ_s. Однако при плохо обусловленных матрицах $\mathbf{m}_s$ рекомендуется использовать в (9.17) дополнительные значения функций, подсчитанные в точках $\Theta_{l_1}, \ldots, \Theta_{l_{k_s}}$. Таким образом, упоминавшееся ранее q равно $\sum\limits_{l=1}^{s} k_l$. Дополнительные точки рекомендуется располагать на направлениях, ортогональных к направлению, определяемому (9.17).

При линейной параметризации метод Ньютона—Гаусса без подсчета производных дает точное решение задачи мнк на первой итерации, если $\rho_1 = 1$.

На наш взгляд, выигрыш в объеме вычислений при переходе от итерационной процедуры (9.18) к более сложным процедурам весьма сомнителен ввиду резкого увеличения сложности расчетов на каждой итерации.

9.5.3. Некоторые замечания о выборе длины шага. Один из главных недостатков изложенного метода состоит в следующем. Как показано ранее, существенным свойством квазиградиентных методов является возможность отыскания такого (может быть, достаточно малого) шага ρ_s в выбранном направлении, который приводит к уменьшению значения минимизируемой функции. Это имеет место всякий раз, когда направление минимизации составляет острый угол с антиградиентом. Выбор же направления в формуле (9.19) не гарантирует нам этого. Поэтому даже при малых $\rho_s > 0$ мы будем не в состоянии уменьшить значение J (Θ). В [242] предлагается попеременно для ρ_s пробовать как положительные, так и отрицательные значения. Например, если задаться коэффициентом редукции $0 < \beta < 1$, то можно положить $\rho = (-\beta)^r$, где $r = 0, 1, \ldots$; ρ_s тогда соответствует первому r, для которого $J(\widehat{\Theta}_{s+1}) < J(\widehat{\Theta}_s)$. Однако и эта процедура иногда может не привести к уменьшению функции (например, в случае, когда направление минимизации ортогонально градиенту). Тогда

следующую точку можно искать методом случайного поиска, в окрестности данной $\widehat{\Theta}_s$.

9.5.4. Разностные аналоги метода Ньютона. В предыдущем пункте отправной точкой при конструировании итерационных процедур служила доступность информации лишь о функциях $f_i(\widehat{\Theta}_s)$. Иногда в регрессионных задачах оказывается сравнительно просто подсчитать производные $\dot{\mathbf{f}}_{is}$. Подчеркнем, что речь идет о непосредственном подсчете величины $\dot{\mathbf{f}}_{is}$, а не их разностных аналогов, которые требуют, по крайней мере $(m+1)$-го вычисления функций $f_i(\widehat{\Theta}_s)$. Обращение к разностным формулам для вычисления $\dot{\mathbf{f}}_{is}$ делает применение изложенной ниже группы методов бессмысленным. Каждый шаг, ими определяемый, будет по трудоемкости близок к нескольким шагам из итерационных процедур, рассмотренных в п. 9.4.1 и 9.5.2.

По-видимому, наиболее известным из рассматриваемой группы методов является метод Дэвидона—Флетчера—Пауэлла. Идея, лежащая в основе данных методов, состоит в отыскании на каждом шаге направлений спуска, близких к направлению метода Ньютона, но без использования матрицы вторых производных.

Матрица $\widetilde{\mathbf{H}}_s$, аппроксимирующая матрицу $\mathbf{H}_s$ из метода Ньютона, может быть, например, выбрана как решение системы

$$\nabla J(\widehat{\Theta}_t)-\nabla J(\widehat{\Theta}_{t-1})=\mathbf{H}^{-1}(\widehat{\Theta}_t-\widehat{\Theta}_{t-1});\ t\leqslant s, \qquad (9.20)$$

где $\mathbf{H}$ — симметричная матрица.

Решение системы (9.20) при $s<m$ неоднозначно, поэтому возможны различные способы конструирования матрицы $\mathbf{H}$ Различные методы рассматриваемой группы отличаются способами конструирования этой матрицы и правилами выбора направлений $\widetilde{\mathbf{H}}_s^{-1}\nabla J(\widehat{\Theta}_s)$, если они не определены однозначно. Матрица $\widetilde{\mathbf{H}}_s$ должна удовлетворять одновременно уравнениям (9.20) при всех $t\leqslant s$.

Приведем способы конструирования матриц $\widetilde{\mathbf{H}}_s$ для некоторых наиболее распространенных методов (см. [108, 115, 240, 211]):

а) *метод сопряженных градиентов*:

$$\widetilde{\mathbf{H}}_{s+1}=\widetilde{\mathbf{H}}_0-\frac{\widetilde{\mathbf{H}}_0\, g_s\, \nabla J(\widehat{\Theta}_s)\,\widetilde{\mathbf{H}}_0'}{\nabla J'(\widehat{\Theta}_s)\,\widetilde{\mathbf{H}}_0\nabla J(\widehat{\Theta}_s)},$$

где $g_s = \nabla J(\widehat{\Theta}_s) - \nabla J(\widehat{\Theta}_{s-1})$, матрицу $\mathbf{H}_0$ часто выбирают единичной или диагональной с элементами $\Theta_{0\alpha}/\nabla J_\alpha(\Theta_0)$, $\alpha = \overline{1, m}$;

б) *метод Дэвидона—Флетчера—Пауэлла*:

$$\widetilde{\mathbf{H}}_s = \widetilde{\mathbf{H}}_{s-1} + \frac{(\widehat{\Theta}_s - \widehat{\Theta}_{s-1})(\widehat{\Theta}_s - \widehat{\Theta}_{s-1})'}{(\widehat{\Theta}_s - \widehat{\Theta}_{s-1})' g_s} - \frac{\widetilde{\mathbf{H}}_{s-1} g_s g_s' \widetilde{\mathbf{H}}_{s-1}}{g_s' \widetilde{\mathbf{H}}_{s-1} g_s},$$

в) *измененный вариант метода Дэвидона—Флетчера—Пауэлла*:

$$\widetilde{\mathbf{H}}_s = \widetilde{\mathbf{H}}_{s-1} + \frac{(\widehat{\Theta}_s - \widehat{\Theta}_{s-1})(\widehat{\Theta}_s - \widehat{\Theta}_{s-1})'}{(\widehat{\Theta}_s - \widehat{\Theta}_{s-1})' g_s} - \frac{\widetilde{\mathbf{H}}_{s-1} g_s (\widehat{\Theta}_s - \widehat{\Theta}_{s-1})'}{(\widehat{\Theta}_s - \widehat{\Theta}_{s-1})' g_s}.$$

При квадратичной функции $J(\Theta)$ (в нашем случае — линейной параметризации) после m шагов матрица $\mathbf{H}$, подсчитываемая любым из методов а)—в), в точности совпадает с матрицей $\mathbf{H}_s$, определенной в п. 9.3. Иными словами, при $\rho_s = 1$ в этом случае точное решение исходной экстремальной задачи будет заведомо получено за m шагов. Если ρ_s выбирается из условия наискорейшего спуска в направлении $-\widetilde{\mathbf{H}}_s \nabla J(\widehat{\Theta}_s)$, то при выполнении не слишком ограничительных условий последовательность $\{\widehat{\Theta}_s\}$ сходится при $s \to \infty$ к $\widehat{\Theta}^*$ со сверхлинейной скоростью независимо от выбора Θ_0.

9.6. Способы нахождения начального приближения

В задаче минимизации функции $J(\Theta)$ первостепенное значение имеет удачный выбор начального приближения Θ_0. Разумеется, невозможно придумать общего правила, которое было бы удовлетворительно для всех случаев, т. е. для всех возможных нелинейных функций $\{f_i\}$. Каждый раз приходится искать свое решение. Ниже предлагается набор некоторых способов нахождения грубых начальных приближений, который на практике может служить отправной точкой поиска удовлетворительных приближений в конкретной задаче.

9.6.1. Поиск на сетке. Особенно эффективен этот метод при небольшом числе *собственно нелинейных параметров*. Часто функции f_i устроены так, что при фиксации значений одних параметров (которые и называем собственно нелинейными)

остальная часть параметров становится линейной. Задаваясь тогда нижней и верхней границей для нелинейных параметров, с некоторым шагом можно устроить перебор вариантов на полученной сетке значений этих собственно нелинейных параметров и выявить ту линейную регрессию, которая приводит к минимальной сумме квадратов.

В качестве примера рассмотрим функцию

$$f(X_i;\ \Theta) = \theta_1 + \theta_2 x_i^{(1)} + \theta_3 e^{\theta_4 x_i^{(2)}}. \tag{9.21}$$

Здесь собственно нелинейным параметром будет θ_4. Допустим, известно, что $\underline{\theta}_4 \leqslant \theta_4 \leqslant \overline{\theta}_4$. Пусть h — шаг для параметра θ_4. Вычислим $K = (\overline{\theta}_4 - \underline{\theta}_4)/h$ линейных регрессий

$$\tilde{f}_k(X_i;\ \Theta) = \theta_1 + \theta_2 x_i^{(1)} + \theta_3 z_{ik},$$

где $z_{ik} = \exp[(\underline{\theta}_4 + hk)x_i^{(2)}]$, $k = 1, \ldots, K$, и найдем для каждой из них минимальную сумму квадратов. Наименьшей из них соответствует оптимальное начальное приближение. В принципе шаг h, от которого зависит «густота» сетки, *может варьироваться*, так что за счет уменьшения величины h значения параметров могут быть найдены с любой точностью.

9.6.2. Преобразование модели. Иногда некоторым преобразованием модель можно свести к линейной или же уменьшить число собственно нелинейных параметров (см. п. 6.2.3). Покажем, как этого можно добиться, на примере логистической кривой

$$f(x_i;\ \theta) = \frac{\theta_1}{1 + \theta_2 \exp(\theta_3 x_i)}.$$

Производя над соответствующими уравнениями регрессии обратное преобразование, получим

$$\frac{1}{y_i} \approx \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1} e^{\theta_3 x_i}.$$

Обозначая $z_i = 1/y_i$, $1/\theta_1 = \theta_1'$, $\theta_2/\theta_1 = \theta_2'$, приходим к новой функции, число линейных параметров которой увеличилось с одного (θ_1) до двух (θ_1' и θ_2'). Оценка для параметра θ_3 в новой модели может быть найдена, например, по предыдущему методу.

Здесь уместно сделать следующее замечание о преобразованиях регрессионных моделей. Следует иметь в виду, что ошибка ε, входившая *аддитивно* в исходное уравнение, после преобразования, вообще говоря, уже не будет аддитивна.

Пользуясь разложением в ряд Тейлора и обозначая преобразование $f_i(\theta)$ через $q(f_i)$, получим, пренебрегая слагаемыми порядка ε_i^3,

$$z_i = q(y_i) = q(f_i + \varepsilon_i) = q(f_i) + \dot{q}(f_i)\varepsilon_i + \frac{1}{2}\ddot{q}(f_i)\varepsilon_i^2.$$

Отсюда следует, что

$$\mathbf{E}z_i = \mathbf{E}q(y_i) = q(f_i) + \frac{1}{2}\ddot{q}(f_i)\sigma^2.$$

Последнее равенство можно взять за основу для анализа задачи с преобразованной моделью.

9.6.3. Разбиение выборки на подвыборки. Для нахождения начального приближения можно разбить всю выборку на m подвыборок (с приблизительно равными объемами), где m — число неизвестных параметров. Для каждой подвыборки найдем средние по y и по X, которые обозначим соответственно $(\overline{y}_j, \overline{X}_j)$, $j = 1, \ldots, m$. Решим систему нелинейных уравнений относительно $\theta_1, \ldots, \theta_m$:

$$\begin{cases} \overline{y}_1 = f(\overline{X}_1;\ \theta_1, \ldots, \theta_m) \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ \overline{y}_m = f(\overline{X}_m;\ \theta_1, \ldots, \theta_m). \end{cases}$$

Решение этой системы и будет являться начальным приближением параметров. Очевидно, для того чтобы данный метод «работал», необходимо, чтобы эта система нелинейных уравнений решалась довольно легко, например аналитически.

9.6.4. Разложение в ряд Тейлора по независимым переменным. Основой итеративной минимизации суммы квадратов является разложение функции регрессии в ряд Тейлора до линейных членов по параметрам. Для нахождения грубого начального приближения иногда бывает полезна процедура аппроксимации регрессии путем разложения ее в ряд Тейлора по независимым переменным X_i. Будем для простоты считать X_i одномерным. Пусть $\overline{x}$ — среднее значение, тогда приближенно

$$f(x_i;\ \Theta) \approx f(\overline{x};\ \Theta) + (x_i - \overline{x}) f_x^{(1)}(\overline{x};\ \Theta) + \ldots + \frac{1}{m!}(x_i - \overline{x})^m f_x^{(m)}(\overline{x};\ \Theta).$$

Обозначим $(x_i - \overline{x})^k = z_{ik}$, $f_x^{(k)}(\overline{x};\ \Theta) = \Phi_k$, таким образом приходим к линейной модели

$$\tilde{f}(Z_i;\ \Phi) = \Phi_0 + \Phi_1 z_{i1} + \ldots + \Phi_m z_{im}.$$

Пусть $\widehat{\Phi}_k$ — мнк-оценки параметров этой линейной регрессии. В качестве начальных приближений примем решение нелинейной системы уравнений относительно $\theta_1, \ldots, \theta_m$:

$$f(\bar{x};\ \Theta) = \widehat{\Phi}_0, \ldots, f_x^{(m)}(\bar{x};\ \Theta) = \widehat{\Phi}_m.$$

Очевидно, этот метод приемлем в том случае, когда последняя система относительно первоначальных параметров решается довольно просто (аналитически).

Покажем использование этого приема на примере нелинейной регрессии (9.21). Для простоты будем считать $\bar{x}^{(2)} = 0$. Тогда

$$e^{\theta_4 x_i^{(2)}} \approx 1 + (x_i^{(2)} - \bar{x}^{(2)})\,\theta_4 + \frac{1}{2}(x_i^{(2)} - \bar{x}^{(2)})^2\,\theta_4^2.$$

Подставляя это разложение в (9.21), получим

$$\tilde{f}(X_i;\ \Theta) = (\theta_1 + \theta_3) + \theta_2 x_i^{(1)} + \theta_3 \theta_4 x_i^{(2)} + \frac{1}{2}\theta_3 \theta_4^2 (x_i^{(2)})^2.$$

Обозначив $\Phi_0 = \theta_1 + \theta_3$, $\Phi_1 = \theta_2$, $\Phi_3 = \theta_3\theta_4$, $\Phi_4 = \theta_3\theta_4^2/2$, $x_i^{(1)} = z_{i1}$, $x_i^{(2)} = z_{i2}$, $(x_i^{(2)})^2 = z_{i3}$, приходим к линейной по Φ модели

$$\tilde{f}(Z_i;\ \Phi) = \Phi_0 + \Phi_1 z_{i1} + \Phi_2 z_{i2} + \Phi_3 z_{i3}.$$

Тогда если $\widehat{\Phi}_j$ — мнк-оценки линейной регрессии, то легко проверить, что начальным приближением для параметров Θ будут $\widehat{\theta}_2 = \widehat{\Phi}_1$, $\widehat{\theta}_4 = 2\widehat{\Phi}_4/\widehat{\Phi}_3$, $\widehat{\theta}_3 = \widehat{\Phi}_3/\widehat{\theta}_4$, $\widehat{\theta}_1 = \widehat{\Phi}_0 - \widehat{\theta}_3$.

В реальности возможна комбинация предложенных способов. Практика показывает, что таким образом можно получить достаточно хорошее начальное приближение для широкого круга нелинейных регрессионных задач.

9.7. Вопросы существования и единственности мнк-оценки

9.7.1. Существование. Запись (9.1), строго говоря, не совсем корректна, так как мнк-оценка может отсутствовать, если априорное множество параметров Γ не является компактом (по предположению f_i считаем непрерывными на Γ). Отсутствие решения в задаче мнк практически приведет к тому, что итеративный процесс минимизации $J(\Theta)$ будет расходиться

и $\|\Theta_s\| \to \infty$ при $s \to \infty$. Естественно, перед тем как решать задачу минимизации, желательно удостовериться, что она корректна. Изложим один подход к решению этой проблемы (более подробно см. [43, 44]). Назовем *нижней границей функции* $J(\Theta)$ *на бесконечности* число

$$\bar{J} = \lim_{r \to \infty} \inf_{\|\Theta\| \geqslant r} J(\Theta).$$

Можно показать, что если существует такое начальное приближение Θ_0, что $J(\Theta_0) < \bar{J}$, то мнк-оценка существует, а множество $S(\Theta_0) = \{\Theta \in R^m : J(\Theta) \leqslant J(\Theta_0)\}$ компактно. Компактность его гарантирует существование хотя бы одной предельной точки последовательности значений параметров, вырабатываемой одним из методов минимизации.

Нелинейная *регрессия имеет бесконечные хвосты*, если при $\|\Theta\| \to \infty$ $|f_i(\Theta)| \to \infty$ для любого $i = 1, 2, \ldots, n$. Наоборот, *регрессия имеет конечный хвост*, если существует такая последовательность параметров $\|\Theta_k\| \to \infty$, $k \to \infty$, что $|f_i(\Theta)| \leqslant M < \infty$ для всех $i = 1, 2, \ldots, n$. Можно показать, что $\bar{J} = +\infty$ тогда и только тогда, когда регрессия имеет бесконечные хвосты. В случае $\bar{J} = \infty$ мнк-оценка всегда существует. Например, в случае логлинейной модели, т. е. когда $f(X_i; \Theta) = \exp(\Theta' X_i)$, регрессия имеет бесконечные хвосты, если векторы $X_1, \ldots, X_n \in R^m$ *разнонаправлены* (для любого $\alpha \in R^m$ существует вектор X_j, для которого $\alpha' X_j < 0$). В то же время можно показать, что если наблюдения $y_i > 0$, то в логлинейной модели мнк-оценка всегда существует.

Оценки снизу для величины $\bar{J}$ в каждом конкретном случае находят аналитически до начала процесса минимизации.

9.7.2. Единственность. Сумма квадратов может быть мультимодальной. Более того, можно доказать, что вероятность мультимодальности $J(\Theta)$ отлична от нуля для любой нелинейной регрессии, не сводящейся к линейной преобразованием в пространстве параметров. Дело усугубляется тем, что все методы минимизации в лучшем случае приводят к локальному минимуму функции. Проверка того, является ли этот минимум глобальным является следующей, возможно, не менее трудоемкой операцией. На практике часто поступают следующим образом. Процесс итераций начинают из другого начального приближения. Тогда, если он сойдется к точке, полученной в первой попытке, можно быть более уверенным в том, что нелинейный минимум является глобальным.

Существуют аналитические методы проверки на достижимость глобального минимума суммы квадратов. Все они пред-

полагают аналитическое исследование поверхности отклика. Иногда при исследовании этой проблемы бывает полезен следующий результат. Пусть найдено выпуклое множество параметров S, на котором гессиан $\partial^2 J/\partial\Theta\partial\Theta'$ положительно определен. Пусть далее найдена оценка снизу M, такая, что $J(\Theta) \geqslant M$ для всех $\Theta \in S$. Тогда, если найденной точке в процессе итераций $\Theta^* \in S$ соответствует локальный минимум со значением $J(\Theta^*) < M$, то $J(\Theta^*)$ — глобальный минимум функции. Например, для логлинейной модели [43] $f(X_i; \Theta) = \exp(\Theta' X_i)$ с положительными наблюдениями y_i множеством S будет

$$S = \{\Theta \in R^m : \Theta' X_i > \ln y_i - \ln 2,\ i = 1, \ldots, n\},$$

а в качестве простейшей оценки снизу можно взять $M = \frac{1}{4} \min y_i^2$. Подобные оценки, как и в случае исследования на существование мнк-оценки, необходимо проводить аналитическими методами.

ВЫВОДЫ

1. При исследовании параметрических моделей регрессии наиболее распространенным типом оптимизируемого (с целью нахождения оценок неизвестных значений параметров регрессии) критерия адекватности модели является *взвешенный* (*или обобщенный*) *критерий наименьших квадратов* (см. (9.1), (9.2)). Следует стремиться к построению таких вычислительных алгоритмов решения оптимизационных задач, которые наряду с решениями этих задач — значениями оценок $\widehat{\Theta}$ неизвестных параметров Θ, — давали бы необходимые характеристики их точности (оценки элементов ковариационных матриц, доверительные области и т. п.).

2. Наибольшее распространение среди методов поиска оценок наименьших квадратов получили алгоритмы итерационного типа, позволяющие на каждой следующей ($(s+1)$-й) итерации получать приближенные значения $\widehat{\Theta}_{s+1}$ искомых оценок параметров, лежащие «ближе» к истинному решению $\widehat{\Theta}^*$ соответствующей оптимизационной задачи, чем значения $\widehat{\Theta}_s$ предыдущей итерации, т. е. $\widehat{\Theta}_{s+1} = \widehat{\Theta}_s + \rho_s \cdot \delta_s$, где s — номер итерации; δ_s— вектор, определяющий направление движения на s-й итерации; ρ_s — длина шага. Если движение осуществляется в направлении под острым углом к антиградиенту оп-

тимизируемой функции, то алгоритм относится к *классу алгоритмов квазиградиентного типа.*

3. Если движение в итерационной процедуре уточнения значений оценок параметров осуществляется непосредственно в направлении антиградиента, то процедуру относят к *алгоритмам градиентного спуска.* Подобные алгоритмы обеспечивают (при определенных ограничениях на минимизируемую функцию) сходимость последовательности $\widehat{\Theta}_s$ со скоростью геометрической прогрессии (линейная сходимость). Из-за того, что реальная скорость сходимости таких алгоритмов резко снижается при приближении $\widehat{\Theta}_s$ к предельному значению $\widehat{\Theta}^*$, градиентный спуск целесообразно применять лишь на начальных этапах минимизации, используя найденные в результате сравнительно небольшого числа итераций величины $\widehat{\Theta}_s$ в качестве начальных приближений для более сложных методов, обладающих большей скоростью сходимости.

4. *В методе Ньютона* значения неизвестных параметров на каждой следующей итерации $\widehat{\Theta}_{s+1}$ находятся из условия минимума квадратичного полинома, аппроксимирующего исходную критериальную функцию в окрестности точки $\widehat{\Theta}_s$. При этом соответствующая процедура будет менее чувствительна к выбору начального приближения (в частности, будет менее подвержена эффекту «раскачки» при его неудачном выборе), если использовать ее вариант с *регулировкой шага.* При определенных условиях метод Ньютона обеспечивает квадратичную скорость сходимости последовательности $\widehat{\Theta}_s$ к $\widehat{\Theta}^*$.

5. Используя *линейную* (*по параметрам*) аппроксимацию исследуемой функции регрессии в окрестности точки $\widehat{\Theta}_s$, можно прийти к модификации метода Ньютона — *методу Ньютона—Гаусса.* Он существенно проще в вычислительном плане, однако бывает слишком чувствительным к эффекту слабой обусловленности используемых в нем матриц $\mathbf{M}_s$. Скорость сходимости этого метода в зависимости от условий, накладываемых на регрессионную функцию и свободные параметры алгоритма, может быть линейной, сверхлинейной или квадратичной.

6. *Существенным недостатком* методов квазиградиентного типа, в том числе метода Ньютона, метода Ньютона—Гаусса и других, является *необходимость подсчета производных* от искомых регрессионных функций на каждой итерации. Основная идея, на которую опираются методы, позволяющие обходиться без подсчета производных, заключается в ис-

пользовании на ($s+1$)-й итерации информации, полученной на предыдущих s итерациях, для построения разумных аппроксимаций для элементов матриц, определяющих выбор направления и шаг движения к решению $\hat{\Theta}^*$.

7. Первостепенное значение для скорости сходимости используемых итерационных процедур решения оптимизационной задачи метода наименьших квадратов имеет *удачный выбор начального приближения* $\hat{\Theta}_0$. Для реализации этого выбора используется ряд приемов: «поиск на сетке» (п. 9.6.1), вспомогательное преобразование (линеаризующее) модели (п. 9.6.2), разбиение имеющейся выборки на подвыборки (п. 9.6.3), разложение регрессионной функции в ряд Тейлора (п. 9.6.4).

8. При вычислительной реализации метода наименьших квадратов в нелинейном (по оцениваемым параметрам Θ) случае приходится исследовать *вопросы существования и единственности решения.* Необходимо помнить, что используемые (в том числе все описанные выше) методы оптимизации приводят в лучшем случае лишь к локальному минимуму критериальной функции. Проверка того, является ли этот минимуму глобальным, является следующей, зачастую не менее трудоемкой, вычислительной операцией.

Глава 10. НЕПАРАМЕТРИЧЕСКАЯ, ЛОКАЛЬНО-ПАРАМЕТРИЧЕСКАЯ И КУСОЧНАЯ АППРОКСИМАЦИЯ РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ

На практике далеко не всегда исходя из профессиональных соображений удается найти аналитический вид регрессионной зависимости. Использование же для ее описания одного из стандартных классов функций может привести к заметной систематической ошибке. Для уменьшения этой опасности прибегают к методам локального (при заданном значении регрессора) оценивания регрессии (§ 10.1—10.2) или же разбивают область возможных значений регрессора на несколько частей и для каждой из них строят свое аналитическое описание регрессионной зависимости (§ 10.3).

Построение простейших непараметрических оценок рассматривается в п. 10.1.1. Их слабое место: недостаточно эффективное использование гладкости регрессии и особенностей геометрического расположения выборочных значений регрессора. Возможны два пути борьбы с этим недостатком: 1) усложнение весовой функции в (10.2) и 2) локальное использо-

вание обычной параметрической регрессии для оценки коэффициентов при первых членах разложения регрессионной кривой (поверхности) в ряд Тейлора в окрестности изучаемой точки. В этой главе принят второй путь как более наглядный и традиционный для статистики.

10.1. Непараметрическое оценивание регрессии

10.1.1. Роль и место непараметрических методов. Непараметрический подход к оцениванию позволяет ослабить два основных требования классической постановки регрессионной задачи. Первое—предположение о том, что $\mathbf{E}\,(y|X)$ как функция X представима в виде $f\,(X;\,\Theta)$, где $f\,(\ldots,\,\ldots)$ — известная функция своих аргументов, а Θ — вектор неизвестных параметров, оцениваемый по выборочным данным, — заменяется на более слабое предположение, что $f\,(X)$ — непрерывная и гладкая функция X. Второе—требование постоянства $\sigma^2\,(X)$— дисперсии случайной погрешности — заменяется на предположение непрерывности $\sigma^2\,(X)$.

В простейшей непараметрической оценке выбирается некоторая непустая окрестность точки X_0 $O\,(X_0)$ и, предполагая, что в этой окрестности $f\,(X)$ приблизительно постоянна, полагаем

$$\hat{f}\,(X_0) = \sum_{X_i \in O\,(X_0)} y_i \Big/ \sum_{X_i \in O\,(X_i)} 1 \tag{10.1}$$

где суммирование в числителе и знаменателе проводится по всем выборочным точкам $X_i \in O\,(X_i)$. Формуле (10.1) можно придать несколько другой вид, удобный для дальнейших обобщений. Введем весовую функцию $w\,(X,\,X_0) = 1$, если $X \in O\,(X_0)$, и равную нулю в противном случае. Тогда (10.1) перепишется в виде

$$\hat{f}\,(X_0) = \sum w\,(X,\,X_0)\,y_i / \sum w\,(X,\,X_0). \tag{10.2}$$

Классическая непараметрическая оценка регрессии получается из (10.2) путем предположения, что

$$w\,(X,\,X_0) \equiv k\,(\|X - X_0\|^{1/2}/b),$$

где $k\,(u)$ — известная функция неотрицательного аргумента, стремящаяся к нулю при $u \to \infty$; b — параметр масштаба, задающий размер окрестности. Обычно полагают

$$k\,(u) = \exp\,\{-u^2/2\} \quad \text{или} \quad k\,(u) = 1/\,(1 + u^2). \tag{10.3}$$

Поскольку в непараметрических оценках $\hat{f}(X_0)$ используется не вся выборка, а только ее часть — совокупность пар (y_i, X_i) с X_i, входящими в окрестность $O(X_0)$, где приближенно верны классические предположения, то *в случае, когда* 1) *классические предположения верны для всей области изменения X и* 2) *параметрическое представление регрессионной зависимости известно исследователю, непараметрические оценки всегда менее эффективны по сравнению с классическими*. Однако они имеют меньшее смещение, когда эти предположения нарушаются. В реальных задачах, выбирая метод оценивания регрессии, следует стремиться сбалансировать обе погрешности: случайную, как правило, уменьшающуюся при расширении объема используемой выборки, и систематическую, растущую при этом.

10.1.2. Примеры. Прежде чем переходить к последовательному изложению непараметрических оценок, приведем два примера их использования, заимствованных из опубликованных работ.

Пример 10.1 [49]. Для анализа производительности труда изучалась зависимость y — выработки (руб.) на одного рабочего в строительно-монтажном тресте от $x^{(1)}$ — объема (млн. руб.) строительно-монтажных работ (СМР); $x^{(2)}$ — рентабельности в процентах к сметной стоимости СМР; $x^{(3)}$ — объема продукции подсобных предприятий в процентах к сметной стоимости СМР; $x^{(4)}$ — ритмичности СМР в течение года по кварталам (%); $x^{(5)}$ — фактической стоимости потребленных материалов в процентах к общей стоимости СМР; $x^{(6)}$ — фондоемкости; $x^{(7)}$ — текучести кадров к среднемесячному числу рабочих. За единицу наблюдения был взят строительно-монтажный трест. Выборка объема $n = 47$.

После проведения классического регрессионного анализа с отсевом незначимых факторов была получена модель

$$y = 5607{,}4 + 63{,}74x^{(1)} + 48{,}65x^{(2)} - 16{,}87x^{(3)} \qquad (10.4)$$

Эта модель легко интерпретировалась с точки зрения экономического содержания. Действительно, $x^{(1)}$ и $x^{(2)}$ являются ведущими аргументами, увеличение которых положительно сказывается на выработке, а $x^{(3)}$ — это производство продукции внутри треста, которое в силу малой мощности предприятий не может быть рентабельным, но без него невозможно строительство. Погрешность аппроксимации в терминах ε — среднего абсолютного относительного отклонения и σ — стандартного отклонения составила для (10.4)

$$\varepsilon_{кл} = \sum \left| \frac{\hat{y}_i - y_i}{y_i} \right| \Big/ n = 0{,}082; \quad \sigma_{кл} = 780.$$

Оценка той же регрессионной зависимости с помощью непараметрической процедуры (10.2) дала заметно лучшее приближение: $\varepsilon_{\text{непар}} = 0{,}051$, $\sigma_{\text{непар}} = 424$. Правда, интерпретация непараметрической модели сложнее. Работа не оканчивается получением оценок значений регрессии в заданных точках, а требуется еще установить, как в среднем меняется регрессионная зависимость при изменении аргументов. Часто этому не уделяется должного внимания.

Пример 10.2 [22, 88]. Рассматривалась модельная одномерная задача, где x — случайная величина, равномерно распределенная на отрезке (0; 10), $f(x) = 10(1 + \exp\{-x/2\})$, $\sigma(x) = 0{,}1 \cdot f(x)$. В качестве меры погрешности i-го метода аппроксимации бралось $\delta_i = \Sigma (f(x_j) - \widehat{f_i}(x_j))^2/n$.

На рис. 10.1 показаны значения δ_i ($i = 0, 1, 2$), соответствующие непараметрическому оцениванию с помощью метода локальной параболической (порядка i) аппроксимации (§ 10.2) с весовой функцией $w(x, x_0) = \exp\{-(x - x_0)^2/2b^2\}$. Параметрическое оценивание с неадекватно предположенной моделью $f_{\text{пар}} = (a + cx)^{-1}$ в обоих случаях ($n = 75$ и $n = 300$) дало значительно большую погрешность $\delta_{\text{пар}} > 1$.

По рисунку видно, что при использовании неадекватной параметрической модели погрешность наибольшая. Локальная параболическая аппроксимация с использованием полинома второй степени лучше, чем традиционно применяемая аппроксимация полиномом нулевой степени. Первая не только дает наименьшую погрешность, но и значительно устойчивее к выбору величины b.

10.1.3. Выбор параметра масштаба b. Это наиболее ответственный момент при использовании непараметрических оценок типа (10.2). Здесь возможны два подхода: 1) выбор единого значения b для всей области изменения X (так обычно поступают на практике) и 2) локальный выбор b в зависимости от того, насколько близко к искомой точке X_0 расположены точки X_i ($i = 1, \ldots, n$) выборки.

В первом случае целесообразно построить как функцию b кривую

$$d^2(b) = n^{-1} \cdot \sum_{i=1}^{n} \left(y_i - \sum_{j \neq i} k(\| X_j - X_i \|^{1/2}/b)\, y_j \Big/ \sum_{j \neq i} k(\| X_j - X_i \|^{1/2}/b) \right)^2, \tag{10.5}$$

задающую асимптотически (при $n \to \infty$) несмещенную оценку величины среднеквадратической погрешности непараметри-

ческой аппроксимации. Подходящее значение b_0 находится из условия, что $d^2(b_0) \leqslant d^2(b)$ для всех $b \neq b_0$.

Локальный выбор b целесообразен, когда X_i расположены в области интересующих нас значений очень неравномерно. В этом случае можно, например, потребовать, чтобы $b(X_0)$ — величина b в точке X_0 — выбиралось из условия

$$b_m(X) = \inf \{b : \text{число}\,(X_i : k(\| X_i - X \|^{1/2}/b) \geqslant 1/2) \geqslant m\}, \quad (10.6)$$

где m — некоторое наперед заданное число. Для нахождения оптимального значения m можно воспользоваться описанной выше процедурой, но с заменой в (10.5) в левой части $d^2(b)$ на $d^2(m)$, а в правой под знаком второй суммы b — на $b_m(X_i)$, где $b_m(X_i)$ определяется (10.6). Так же как b, $m_{\text{опт}}$ находится из условия минимизации $d^2(m)$.

10.1.4. Более эффективное использование гладкости $f(X)$. Если регрессионная поверхность достаточно гладкая и в окрестности X_0 может приближенно считаться линейной, т. е.

$$f(X) = f(X_0) + \Theta'(X_0)(X - X_0) + O(\| X - X_0 \|), \quad (10.7)$$

где $\Theta(X_0) = (\theta^{(1)}, \ldots, \theta^{(p)})'$ — неизвестный вектор, зависящий только от X_0, то для оценки $f(X_0)$ вместо (10.2) можно воспользоваться оценкой

$$\hat{f}(X_0) = \begin{vmatrix} z_0 & u_{01} & \ldots & u_{0p} \\ z_1 & u_{11} & \ldots & u_{1p} \\ \cdot & \cdot & \cdot & \cdot \\ z_p & u_{p1} & \ldots & u_{pp} \end{vmatrix} : \begin{vmatrix} u_{00} & u_{01} & \ldots & u_{0p} \\ u_{10} & u_{11} & \ldots & u_{1p} \\ \cdot & \cdot & \cdot & \cdot \\ u_{p0} & u_{p1} & \ldots & u_{pp} \end{vmatrix}, \quad (10.8)$$

где $z_0 = \sum y_i w(X_i, X_0)$; $z_j = \sum y_i (x_i^{(j)} - x_0^{(j)}) w(X_i, X_0)$ $(j \neq 0)$; $u_{j0} = u_{0j} = \sum_i (x_i^{(j)} - x_0^{(j)}) w(X_i, X_0)$;

$$u_{kj} = \sum_i (x_i^{(k)} - x_0^{(k)})(x_i^{(j)} - x_0^{(j)}) w(X_i, X_0)$$

для $k, j \neq 0$. Для обоснования (10.8) заметим, что предложенная точечная оценка есть обычная мнк-оценка постоянного члена в случае линейной поверхности регрессии, когда в (10.7) отброшены члены, обозначенные $O(\|X - X_0\|)$. В примере 10.2 мы видели, что эта оценка (соответственно δ_1) может быть значительно лучше оценки (10.2) (соответственно δ_0). Она специально рассчитана на случай, когда X_i не заполняют равномерно пространство возможных значений X.

Использование оценки (10.8) вместо (10.2) открывает возможность содержательной интерпретации регрессионной за-

висимости, на необходимость чего было обращено внимание в примере 10.1. Для этого достаточно наряду с оценкой $\widehat{f}(X_0)$ построить $\widehat{\Theta}(X_0)$ — оценку вектора $\Theta(X_0)$ в (10.7)

$$\widehat{\theta}^{(i)} = \begin{vmatrix} u_{00} & \dots & u_{0,(i-1)} & z_0 & u_{0,(i+1)} & \dots & u_{0p} \\ u_{10} & \dots & u_{1,(i-1)} & z_1 & u_{1,(i+1)} & \dots & u_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ u_{p0} & \dots & u_{p,(i-1)} & z_p & u_{p,(i+1)} & \dots & u_{pp} \end{vmatrix} : \begin{vmatrix} u_{00} & \dots & u_{0p} \\ u_{10} & \dots & u_{1p} \\ \cdot & \cdot & \cdot \\ u_{p0} & \dots & u_{pp} \end{vmatrix} . \quad (10.9)$$

Сравнение $\widehat{\Theta}(X)$ для разных значений X дает возможность оценить, как, насколько меняется влияние изучаемых факторов на регрессию в разных областях пространства возможных значений X.

При использовании формул (10.8), (10.9) выбор величины b можно производить так же, как указано в предыдущем пункте.

10.2. Локальная параметрическая аппроксимация регрессии в одномерном случае

Основная цель этого параграфа — дать в краткой форме теоретическое объяснение тем фактам, которые были выявлены путем моделирования в примере 10.2. Это целесообразно сделать потому, что локальная параметрическая аппроксимация не нашла еще достаточного отражения в теоретических исследованиях и мало используется в практических работах.

10.2.1. Основная формула для оценки. Из разложения $f(x)$ в ряд Тейлора в $O(x_0)$ — окрестности $x = x_0$ до членов порядка l

$$f(x) = f(x_0) + \sum_{1}^{l} f^{(i)}(x - x_0)^i + o(|x - x_0|^l) \quad (10.10)$$

и уравнений мнк (см. гл. 7) получаем оценку

$$\widehat{f}(x_0) = \begin{vmatrix} z_0 & u_1 & \dots & u_k \\ z_1 & u_2 & \dots & u_{k+1} \\ \cdot & \cdot & \cdot & \cdot \\ z_k & u_{k+1} & \dots & u_{2k} \end{vmatrix} : \begin{vmatrix} u_0 & \dots & u_k \\ u_1 & \dots & u_{k+1} \\ \cdot & \cdot & \cdot \\ u_k & \dots & u_{2k} \end{vmatrix}, \quad (10.11)$$

где $z_j = \Sigma y_i v_i^j w(v_i)$; $u_j = \Sigma v_i^j w(v_i)$; $v = x - x_0$; $w(v) = 1$, когда $x_0 + v \in O(x_0)$ и равно нулю в противном случае. Если рассматривать только пары точек (y_i, x_i), где $x_i \in O(x_0)$, то

$\widehat{f}(x_0)$ — это просто мнк-оценка постоянного члена в (10.10), когда пренебрегают вкладом $o(|v|^l)$. На практике за $w(v)$ обычно берутся функции типа (10.3). Однако в теоретическом исследовании мы выберем

$$w(v) = \begin{cases} 1, \text{ если } |v| \leqslant b; \\ 0, \text{ если } |v| > b. \end{cases} \quad (10.12)$$

10.2.2. Асимптотическая оценка точности приближения $f(x_0)$. Пусть в окрестности x_0 f непрерывна и имеет l первых непрерывных производных $f^{(i)}(x)$, причем $f^{(l)}(x)$ удовлетворяет условию Липшица порядка $0 < \alpha \leqslant 1$, т. е. для некоторого $c < \infty$ $|f^{(l)}(x+v) - f^{(l)}(x)| \leqslant c|v|^\alpha$. Предположим далее, что точки $x_1, \ldots, x_n$ распределены независимо друг от друга с плотностью $p(x)$. Пусть далее $p(x)$ и $\sigma(x)$ — стандартное отклонение погрешности в точке x — непрерывны в окрестности $x = x_0$ и $p(x_0) > 0$.

При сделанных предположениях при росте объема выборки $(n \to \infty)$ и $b \to 0$ имеем, что дисперсия случайной ошибки есть величина порядка $1/nb$, а квадрат систематического смещения $\widehat{f_l}(x_0)$ не превосходит $b^{2l+2\alpha}$. Уравновешивая обе погрешности, получаем $b = n^{-1/(1+2l+2\alpha)}$, и среднеквадратическое отклонение оценки $\widehat{f_l}(x_0)$ будет величиной порядка $n^{-(l+\alpha)/(1+2l+2\alpha)}$. Заметим, что для гладких функций f b убывает очень медленно.

10.2.3. Сравнение $\widehat{f_0}$ и $\widehat{f_1}$. Пусть в некоторой окрестности $x = x_0$ регрессия $f(x)$ линейна, $f(x) = f(x_0) + \theta(x - x_0)$, а $\sigma(x)$ и $p(x)$ постоянны. Обозначим m число точек $x_i \in [x_0 - b, x_0 + b]$. Пусть далее число наблюдений $n \to \infty$, $w(v)$ определяется (10.12), а $b \to 0$, тогда обе оценки $\widehat{f_0}$ и $\widehat{f_1}$ с некоторого момента несмещены и

$$\mathbf{D}(\widehat{f_0} \mid m) = (\sigma^2 + \theta^2 b^2/3)/m; \quad (10.13)$$

$$\mathbf{D}(\widehat{f_1} \mid m) = \sigma^2(1 + m^{-1} + O(m^{-2}))/m. \quad (10.14)$$

Формулы (10.13) и (10.14) асимптотически эквивалентны. Однако члены второго порядка малости в них различны. При $b = n^{-\delta}$, где $\delta < 1/3$, и $\theta \neq 0$ члены второго порядка малости в (10.14) асимптотически меньше, чём в (10.13). Если $p \neq \text{const}$, то даже при линейной функции f оценка $\widehat{f_0}$ смещена если только $\theta \neq 0$. Как показывает модельный пример 10.2, вклад выборочных флуктуаций x даже при значениях n по-

рядка нескольких сотен заметен. Тем более неравномерное распределение точек X_i должно сказываться в многомерном случае, что служит еще одним аргументом в пользу оценок (10.8) по сравнению с оценками (10.2).

10.2.4. Изучение дисперсии оценок $\widehat{f}_l$ ($l \geqslant 2$). По рис. 10.1 видно, что при малых значениях b $\delta_2 > \delta_0$. Для того чтобы изучить этот эффект, сделаем дополнительное упрощающее предположение, что x_i лежат на равном расстоянии друг от друга. Предположим далее, что все оценки $\widehat{f}_j$ несмещены для

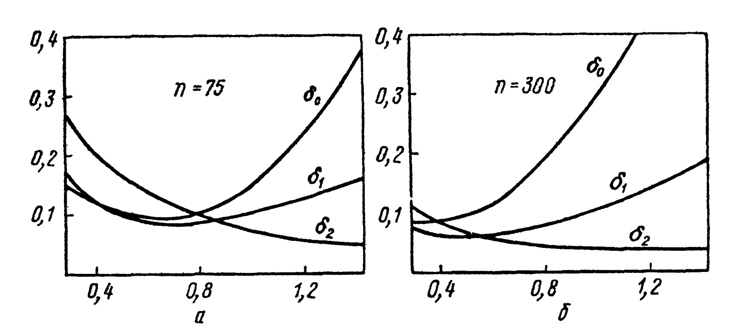


Рис. 10.1. Зависимость погрешности локальной параболической аппроксимации от величины b для выборки объема: а) $n = 75$; б) $n = 300$

$|x - x_0| < b$, и при фиксированном b и $m \to \infty$ оценим ряд отношений $\mathbf{D}\widehat{f}_0 : \mathbf{D}\widehat{f}_1 : \mathbf{D}\widehat{f}_2 : \ldots$ Для этого удобно использовать полиномы Чебышева, ортогональные на $|x - x_0| \leqslant b$ [77]. Обозначим $\varphi_j(v)$ — полином j-го порядка, где $v = x - x_0$; c_j — коэффициент перед φ_j в разложении $f(x_0 + v)$, и $\widehat{c}_j$ — оценку c_j. Очевидно, $\widehat{f}_l = \sum_{j=0}^{l} \widehat{c}_j \varphi_j(0)$. Откуда

$$\mathbf{D}\widehat{f}_l = \sum_{j=0}^{l} \mathbf{D}\widehat{c}_j\, \varphi_j^2(0) \equiv \sum_{j=0}^{l} d_j, \tag{10.15}$$

поскольку оценки $\widehat{c}_j$ независимы. Используя явный вид полиномов Чебышева [77], получаем

$$d_0 = \sigma^2/m;\; d_1 = 0;\; d_2 = 5d_0/4;\; d_3 = 0;\; \ldots$$

$$d_{2l} = d_{2l-2}\,\frac{(2l-1)^2(4l+1)}{(2l)^2(4l-3)};\; d_{2l+1} = 0;\; \ldots,$$

или

$\mathbf{D}\widehat{f}_0 : \mathbf{D}\widehat{f}_2 : \mathbf{D}\widehat{f}_4 = 1 : 2{,}25 : 3{,}51 \ldots$

Из приведенного рассуждения можно сделать вывод, что при $p(x)$, непрерывном в окрестности $x = x_0$, $p(x_0) \neq 0$ и достаточно малом b переход от $\widehat{f}_{2l}$ к $\widehat{f}_{2l+1}$ не вызывает заметного роста дисперсии оценки, в то время как при переходе к $\widehat{f}_{2(l+1)}$ дисперсия оценки увеличивается приблизительно на 1,25 $\mathbf{D}\widehat{f}_0$. Таким образом, лучше использовать оценки нечетного порядка ($\widehat{f}_1$, $\widehat{f}_3$, и т. д.).

Напомним, что контроль среднеквадратического отклонения, учитывающего и случайную, и систематическую составляющие ошибки, целесообразно проводить с помощью кривой (10.5).

10.3. Кусочно-параметрическая (сплайновая) техника аппроксимации регрессионных зависимостей

В последние два десятилетия в вычислительной математике и в инженерной практике широкое распространение получили функции, называемые сплайнами. Этот термин произошел от английского spline, означающего упругую и гибкую металлическую линейку, использовавшуюся для проведения гладкой кривой, проходящей через заданные точки. Одномерный сплайн степени l представляет собой функцию, непрерывную вместе со своими $(l-1)$-ыми производными, у которой производная l-го порядка постоянна на интервалах между заданными точками, называемыми *узлами*. Сплайн l-й степени можно представить состоящим из гладко (до $(l-1)$-го порядка) склеенных в узлах полиномов l-й степени.

Сплайны сравнительно мало известны прикладным статистикам. Вместе с тем, по мнению ряда авторов [54, 114], они являются наиболее удачными аппроксимирующими функциями для приложений. Дело здесь в том, что поведение функции, выражающей физические взаимоотношения, в одной области пространства может быть полностью не связанным с ее поведением в другой области. Полиномы наряду с большинством других математических функций обладают как раз обратным свойством. Их поведение в малой области однозначно определяет поведение в любой другой точке. Сплайны, поскольку они определяются кусочно, лишены этого недостатка, и для

$l \geqslant 3$ они прекрасно представляют гладкие кривые физического мира. В последние годы сплайны стали широко использоваться при аппроксимации регрессионных зависимостей.

Основная часть этого параграфа посвящена одномерным сплайнам.

10.3.1. Определение одномерных сплайнов. Пусть на отрезке $[a, b]$ выделено m точек $u_1, \ldots, u_m$, которые мы будем называть узлами, $P_i(m)$ $(i = 1, \ldots, m+1)$ — полиномы степени l, удовлетворяющие условию

$$P_i^{(k)}(u_i) = P_{i+1}^{(k)}(u_i),\ k = 0, \ldots, l-1;\ i = 1, \ldots, m, \tag{10.16}$$

где $P^{(k)}$ означает производную (по x) k-го порядка от P. Тогда

$$S_l(x) = P_i(x),\ u_{i-1} \leqslant x \leqslant u_i\ (u_0 = a,\ u_{m+1} = b) \tag{10.17}$$

называют сплайном l-го порядка с m узлами. Таким образом, если исследователь хочет использовать сплайны, он должен определить:

а) l — порядок сплайна, т. е. степень полиномов $P_i(x)$;

б) m — число узлов, равное числу различных полиномов без одного;

в) положение узлов u_i $(i = 1, \ldots, m)$ на $[a, b]$;

г) $m + l + 1$ свободных коэффициентов сплайн-функции (каждый полином имеет $(l + 1)$ коэффициент и каждое условие гладкости (10.16) накладывает l связей, откуда число свободных параметров равно: $(m + 1)(l + 1) - ml = m + l + 1$).

Полиномы P_i могут быть представлены в одной из двух форм:

$$P_i(x) = \sum_{k=0}^{l} a_i^{(k)} (x - u_{i-1})^k$$

или

$$P_i(x) = \sum_{k=0}^{l} b_i^{(k)} (x - u_i)^k.$$

Отсюда для сплайна на отрезке $[a, b]$ получаем представление [54]

$$S_l(x) = P_1(x) + \sum_{i=1}^{m} c_i (x - u_i)_+^l, \tag{10.18}$$

где $c_i = a_{i+1}^{(l)} - b_i^{(l)}$, а знак $_+$ справа снизу скобок означает усечение

$$v_+ = \begin{cases} v \text{ при } v > 0; \\ 0 \text{ при } v \leqslant 0. \end{cases}$$

В формулу (10.18) линейно входят только $m + l + 1$ неизвестных коэффициентов, поэтому в принципе она могла бы быть использована в методе наименьших квадратов или в какой-либо другой подходящей процедуре оценивания. Однако с вычислительной точки зрения иметь дело со степенными функциями не удобно и желательно использовать другое представление $S_l(x)$ через так называемые базисные сплайны.

Для этого введем дополнительно l узлов слева $u_{-l} < \ldots \ldots < u_{-1} < a$ и l узлов справа $b < u_{m+2} < \ldots < u_{m+l+1}$. Например, можно положить

$$u_k = \begin{cases} u_0 + k(u_1 - u_0) & \text{при } k < 0; \\ u_{m+1} + k(u_{m+1} - u_m) & \text{при } k > m+1. \end{cases}$$

Определим теперь базисные сплайны (B-сплайны) как (ниже знак i справа сверху B — индекс, а не знак возведения в степень)

$$B_l^i(x) = \sum_{k=i}^{i+l+1} \left[(x - u_k)_+^l \Big/ \prod_{\substack{j=i \\ j \neq k}}^{i+l+1} (u_k - u_j) \right]. \qquad (10.19)$$

B_l^i-сплайны обладают [54] следующим свойством:

$$B_l^i(x) \begin{cases} > 0 \text{ для } x \in (u_i, u_{i+l+1}); \\ \equiv 0 \text{ для } x \notin (u_i, u_{i+l+1}). \end{cases} \qquad (10.20)$$

Сплайн-функция $S_l(x)$ однозначно представляется в виде

$$S_l(x) = \sum_{i=-l}^{m} \theta_i B_l^i(x). \qquad (10.21)$$

Для равноотстоящих на единицу масштаба узлов базисные сплайны B_0, B_1, B_2, B_3 показаны на рис. 10.2.

10.3.2. Выбор порядка сплайна, числа и положения узлов. Это важная и ответственная задача, по своей методической роли эквивалентная выбору класса аппроксимирующих функций в обычном регрессионном анализе. От ее успешного решения существенно зависит, удастся ли при анализе данных использовать все преимущества, представляемые сплайнами, или нет. Здесь трудно дать рекомендации, верные для всех практических задач. Однако, следуя [258], мы попытаемся вы-

сказать некоторые общие соображения для случая, когда наблюдений относительно немного или они распределены крайне неравномерно вдоль оси регрессора. В этих условиях желательно:

1) использовать сплайны 3-го порядка;

2) вводить настолько мало узлов, насколько это возможно. На интервал между узлами иметь не менее 4 или 5 наблюдений. Это правило вызвано тем, что проблема «сверхподгонки» представляет для сплайнов реальную опасность;

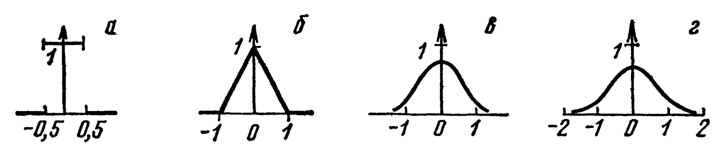


Рис. 10.2. Базисные сплайны с равноотстоящими узлами:
а) B_0; б) B_1; в) B_2; г) B_3

3) иметь не более одной экстремальной точки (максимум или минимум) на интервале. Желательно, чтобы эта точка приходилась на центр соответствующего интервала, а точки перегиба линии регрессии были в окрестности узлов.

Другой подход к выбору узлов можно найти в [197, 245].

10.3.3. Оценка параметров и проверка гипотез. Если при фиксированном порядке сплайна l и заданном положении узлов верны классические предположения регрессионного анализа, т. е. $y_i = S_l(x_i) + \xi_i$, причем ξ_i взаимно независимы, не зависят от x_i и $\xi_i \in N(0; \sigma^2)$, где σ — неизвестная постоянная, то оценка параметров $\Theta = (\theta_{-l}, \ldots, \theta_m)'$ проводится с помощью мнк (см. § 7.1). Обозначим $\widehat{\Theta}$ — мнк-оценку $\widehat{\Theta} = (\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}Y$, где $Y = (y_1, \ldots, y_n)'$, $\mathbf{V}$ — $n \times (m + l + 1)$-матрица с элементами $v_{ij} = B_l^{j-l-1}(x_i)$. В сделанных предположениях $\widehat{\Theta}$ имеет нормальное распределение со средним Θ и ковариационной матрицей

$$\mathbf{C} = \mathbf{E}(\widehat{\Theta} - \Theta)\cdot(\widehat{\Theta} - \Theta)' = \sigma^2(\mathbf{V}'\mathbf{V})^{-1}.$$

Оценка для σ^2 строится стандартным образом как

$$s^2 = \sum (y_i - \widehat{S}_l(x_i))^2/(n - m - l - 1),$$

где $\widehat{S}_l(m) = \sum_{i=-l}^{m} \widehat{\theta}_i B_l^i(x)$ (§ 7.1 или [14, п. 8. 6. 3]).

Из (10.20) следует, что в матрице $\mathbf{C} = \|c_{ij}\|$ для $|i - j| \geqslant l + 1$ $c_{ij} = 0$, т. е. матрица имеет $(2l + 1)$-диагональный вид.

Основные гипотезы, связанные с кубическими сплайн-функциями ($l = 3$). Рассмотрим гипотезы о поведении сплайна между узлами.

Гипотеза 1: *между узлами u_{j-1}, u_j кубический сплайн является квадратическим.* Используя точки как знак дифференцирования $S_3(x)$ по x, эту гипотезу можно выразить как

$$\dddot{S}_3(u_{j-1}) - \dddot{S}_3(u_j) = 0, \tag{10.22}$$

или в терминах B-сплайнов:

$$\sum_{i=-3}^{m} \theta_i(\delta_i - \gamma_i) = 0, \quad \text{или} \quad \left[\sum_{i=-3}^{m} \theta_i(\delta_i - \gamma_i)\right]^2 = 0,$$

где $\delta_i = \dddot{B}_3^i(u_{j-1})$, $\gamma_i = \dddot{B}_3^i(u_j)$ — известные постоянные, зависящие только от расположения узлов u_j, $j = -3, \ldots, m + 4$. Последняя формула может быть использована для построения F-критерия для проверки (10.22). Обозначим $D = (\delta_{-3} - \gamma_{-3}, \ldots, \delta_m - \gamma_m)'$, тогда в случае, когда гипотеза (10.22) верна, величина $FD'\hat{\Theta}\hat{\Theta}'D/D'(\mathbf{V}'\mathbf{V})^{-1}Ds^2$ имеет $F(1, n-m-4)$-распределение.

Гипотеза 2: *на отрезке между узлами u_{j-1} и u_j кубический сплайн линеен.* В использованных выше обозначениях эту гипотезу можно представить как

$$\ddot{S}_3(u_{j-1}) = \ddot{S}_3(u_j) = 0 \tag{10.23}$$

или

$$(\Sigma\, \theta_i\, \delta_i)^2 + (\Sigma\, \theta_i\, \gamma_i)^2 = 0.$$

Последняя форма удобна для построения F-критерия. Обозначим $\Delta = (\delta_{-3}, \ldots, \delta_m)'$, $\Gamma = (\gamma_{-3}, \ldots, \gamma_m)'$ и найдем распределение двумерного вектора $Z = (\Delta'\hat{\Theta}, \Gamma'\hat{\Theta})'$. Вектор Z нормален, имеет (2×2)-ковариационную матрицу $\sigma^2 (\Delta, \Gamma)' \times (\mathbf{V}'\mathbf{V})^{-1} (\Delta, \Gamma)$ и в случае, когда (10.23) имеет место, нулевые средние. Поэтому для проверки гипотезы (10.23) может быть предложен критерий

$$F = Z'[(\Delta, \Gamma)'(\mathbf{V}'\mathbf{V})^{-1}(\Delta, \Gamma)]^{-1} Z/2s^2,$$

где F имеет $F(2, n-m-4)$-распределение.

10.3.4. Билинейные сплайны. Наряду с одномерными сплайн-функциями в приложениях, особенно в экономике [114], получили распространение простейшие сплайны, задаваемые с помощью прямоугольной решетки. Внутри каждого из прямоугольников решетки они представляют билинейную функцию своих аргументов (x, y)

$$z = a + bx + cy + dxy, \qquad (10.24)$$

согласованную таким образом, чтобы z было непрерывной функцией (x, y) при переходе от одного прямоугольника к другому. Пусть на оси x выделено $I + 1$ точек $\Delta_u = \{u_0, u_1, \ldots, u_I\}$ и на оси y $J + 1$ точек $\Delta_v = \{v_0, v_1, \ldots, v_J\}$.

Пусть далее

$$x_i = \begin{cases} x - u_{i-1}, & \text{если } x > u_{i-1} \ (i = 1, \ldots, I); \\ 0 &, \text{если } x \leqslant u_{i-1}; \end{cases}$$

$$y_j = \begin{cases} y - v_{j-1}, & \text{если } y > v_{j-1} \ (j = 1, \ldots, J); \\ 0 &, \text{если } x \leqslant v_{j-1}; \end{cases}$$

$$w_{ij} = x_i y_j, \ i = 1, \ldots, I, \ j = 1, \ldots, J;$$

$$X = (x_1, \ldots, x_I)'; \ Y = (y_1, \ldots, y_J)'; \ W_i = (w_{i1}, \ldots, w_{iJ})';$$

$$\mathbf{W} = (W'_1, \ldots, W'_I)'.$$

Билинейным сплайном на Δ_u, Δ_v называется функция вида

$$z = \theta + \mathrm{A}' X + \mathrm{B}' Y + \Gamma' \mathbf{W}, \qquad (10.25)$$

где $\mathrm{A} = (\alpha_1, \ldots, \alpha_I)'$, $\mathrm{B} = (\beta_1, \ldots, \beta_J)'$, $\Gamma = (\gamma_{11}, \ldots, \gamma_{IJ})'$ — вектор-столбцы параметров размерности соответственно I, J, IJ. Нетрудно видеть, что билинейный сплайн непрерывен и зависит от $1 + I + J + IJ$ параметров. Заметим, что если бы не было условий согласования значений функций (10.24) на решетке, то сплайн зависел бы от $4IJ$ параметров.

Пусть $u_{i-1} \leqslant x \leqslant u_i$ и $v_{j-1} \leqslant y \leqslant v_j$. Представим z в локальных координатах $(x - u_{i-1})$ и $(y - v_{j-1})$:

$$z = a_{ij} + b_{ij}(x - u_{i-1}) + c_{ij}(y - v_{j-1}) + \\ + d_{ij}(x - u_{i-1})(y - v_{j-1}). \qquad (10.26)$$

Эта форма представления удобна для содержательной интерпретации двумерных сплайнов.

Существует простая связь между представлениями (10.25) и (10.26) сплайна. Пусть $\mathbf{A} = \|a_{ij}\|$, $\mathbf{B} = \|b_{ij}\|$, $\mathbf{C} = \|c_{ij}\|$, $\mathbf{D} = \|d_{ij}\|$,

$$\underset{((I+1)\times(J+1))}{\Pi} = \begin{bmatrix} \theta & \beta_1 & \ldots & \beta_J \\ \alpha_1 & \gamma_{11} & \ldots & \gamma_{1J} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_I & \gamma_{I1} & \ldots & \gamma_{IJ} \end{bmatrix}$$

$$\underset{(2I\times(I+1))}{\mathbf{U}} = \begin{bmatrix} 1 & u_0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & u_1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & u_2 & u_2-u_1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & u_{I-1} & u_{I-1}-u_1 & u_{I-1}-u_2 & \ldots & u_{I-1}-u_{I-2} & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 1 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 1 & 1 & 1 & \ldots & 1 & 1 \end{bmatrix}$$

и матрица $\mathbf{V}$ размерности $2J\times(J+1)$ определена аналогично $\mathbf{U}$ с заменой элементов u_i на v_j, тогда [114]

$$\mathbf{U}\Pi\mathbf{V}' = \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{B} & \mathbf{D} \end{bmatrix}$$

ВЫВОДЫ

1. *Непараметрический подход* к оцениванию регрессии *позволяет ослабить ряд основных предположений регрессионного анализа*: 1) требование априорного знания с точностью до неизвестных значений параметров аналитического вида регрессионной зависимости $\mathbf{E}\,(y|X)$ и 2) требование постоянства (для всех значений регрессора) дисперсии случайной погрешности $\mathbf{E}\,((y - \mathbf{E}\,(y|X))^2|X)$. В простейшем случае $\widehat{m}$ — непараметрическая оценка $\mathbf{E}\,(y|X_0)$ — строится следующим образом: выбирается $O\,(X_0)$ — некоторая окрестность X_0; выделяются все пары наблюдений (y_i, X_i), такие, что $X_i \in O\,(X_0)$; пусть таких пар будет k и $u = \Sigma y_i$, где суммирование проводится по всем выделенным парам, тогда $\widehat{m} = u/k$. При выборе

диаметра $O(X_0)$ приходится уравновешивать два источника погрешности: при увеличении диаметра растет отклонение $\mathbf{E}(y|X)$ от $\mathbf{E}(y|X_0)$, а при его уменьшении падает эффективность оценивания.

2. Использование традиционных регрессионных моделей (линейных при многомерном X и параболических в одномерном случае) в применении к относительно большим подобластям изменения регрессора позволяет сочетать простоту расчетов, свойственную классическим моделям регрессии, с эффективным использованием выборочной информации. Эти методы получили название *локально параметрических*.

3. В последние годы для описания регрессионной зависимости стали широко использоваться сплайны. *Сплайном* называют конечную совокупность гладко склеенных между собой полиномов, каждый из которых определен на своей подобласти изменений регрессора. Подобно локально параметрическим методам оценивания сплайны позволяют удачно сочетать достоинства локальных методов (уменьшение смещения оценки) с высокой эффективностью параметрических процедур оценивания.

Глава 11. ИССЛЕДОВАНИЕ ТОЧНОСТИ СТАТИСТИЧЕСКИХ ВЫВОДОВ В РЕГРЕССИОННОМ АНАЛИЗЕ

После реализации этапов $1 \sim 6$, связанных с построением оценки (аппроксимации) $\widehat{f}(X)$ для искомой регрессионной зависимости $f(X) = \mathbf{E}(\eta|\xi = X)$ (см. § В.6), исследователю необходимо ответить на вопросы: *какова точность* полученной им оценки (аппроксимации) $\widehat{f}(X)$ и, в частности, как определить ту гарантированную (с заданной доверительной вероятностью P) величину погрешности, за пределы которой мы не выйдем, восстанавливая неизвестные нам значения параметров θ_k, истинной функции регрессии $f(X)$ или анализируемого результирующего показателя $\eta(X) = (\eta|\xi = X)$ по значениям оценок соответственно $\widehat{\theta}_k$ и $\widehat{f}(X)$.

Достаточно исчерпывающие и теоретически обоснованные ответы на эти вопросы мы в состоянии дать лишь в рамках схемы, постулирующей, что: а) выбор класса $\mathbf{F}$ допустимых решений (т. е. выбор общего параметрического вида функции регрессии $f(X)$) *осуществлен удачно*, а именно: $f(X) \in \mathbf{F}$; б) имеется априорная информация о вероятностной природе (например, о типе закона распределения) регрессионных ос-

татков ε в моделях вида (В.14), (В.16) и (В.21). Будем называть эту схему *идеализированной*. Если же у нас нет оснований рассчитывать на выполнение постулатов а) и б) (что, к сожалению, и бывает в большинстве реальных ситуаций, а потому будем называть эту схему *реалистической*), то получить сколько-нибудь законченные и теоретически обоснованные результаты по оценке точности статистических выводов в регрессионном анализе не удается. В этом случае можно предложить лишь некоторые *полуэвристические приемы* и рекомендации, нацеленные на приближенное решение данной задачи.

11.1 Линейный (относительно оцениваемых параметров) нормальный вариант идеализированной схемы регрессионной зависимости

В данном параграфе рассматривается регрессионная модель зависимости случайного результирующего показателя η от неслучайных объясняющих переменных $\xi=(\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(p)})'$ вида

$$\eta = \sum_{k=0}^{m} \theta_k \psi_k (\xi) + \varepsilon, \qquad (11.1)$$

где $\{\psi_k (\xi)\}_{k=\overline{0,m}}$ — система *известных* (базисных) функций (в частном случае $\psi_0 (\xi) \equiv 1$, $\psi_k (\xi) = \xi^{(k)}$ для $k = 1, 2, \ldots, p$), $\Theta = (\theta_0, \theta_1, \ldots, \theta_m)'$ — неизвестные (подлежащие оцениванию) параметры, а остаточная случайная компонента ε подчиняется *нормальному* закону распределения со средним значением $\mathbf{E}\varepsilon = 0$, и с дисперсией (вообще говоря, неизвестной) $\mathbf{D}\varepsilon = \sigma^2$, т. е.

$$\varepsilon \in N(0, \sigma^2). \qquad (11.2)$$

Отсюда, в частности, следует, что истинная функция регрессии $f(X) = \mathbf{E}(\eta | \xi = X)$ имеет вид

$$f(X) = f(X;\ \Theta) = \theta_0 \psi_0 (X) + \theta_1 \psi_1 (X) + \ldots \theta_m \psi_m (X), \qquad (11.3)$$

т. е. является *линейно зависящей* от неизвестных параметров Θ (форма ее зависимости от X определяется выбором системы базисных функций $\{\psi_k (X)\}_{k=\overline{0,m}}$.)

Соотношение (11.1) определяет связи между имеющимися наблюдениями $\{(X_i y_i)\}_{i=\overline{1,n}}$ вида

$$Y = \mathbf{X}\,\Theta + \varepsilon, \tag{11.4}$$

где $Y = (y_1, y_2, \ldots, y_n)'$ — вектор-столбец наблюденных значений результирующего показателя, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ — вектор-столбец ненаблюдаемых регрессионных остатков, а

$$\mathbf{X} = \begin{pmatrix} \psi_0(X_1)\, \psi_1(X_1) \ldots \psi_m(X_1) \\ \psi_0(X_2)\, \psi_1(X_2) \ldots \psi_m(X_2) \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ \psi_0(X_n)\, \psi_1(X_n) \ldots \psi_m(X_n) \end{pmatrix} —$$

матрица плана, т. е. матрица значений базисных функций в наблюденных точках предикторной переменной. При этом постулируется, что нормально распределенные регрессионные остатки $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ *взаимно некоррелированы*, т. е. что их ковариационная матрица $\mathbf{V} = (\mathbf{E}(\varepsilon_i \varepsilon_j))_{i,j=\overline{1,n}}$ имеет вид

$$\mathbf{V} = \sigma^2 \cdot \mathbf{I}_n, \tag{11.5}$$

где $\mathbf{I}_n$, как обычно, единичная матрица размерности $n \times n$.

Из (11.2) и (11.5) имеем

$$\boldsymbol{\varepsilon} \in N_n(0, \sigma^2 \cdot \mathbf{I}_n). \tag{11.2$'$}$$

Предполагается также, что этап выбора общего параметрического вида искомой зависимости (этап 4, см. § В.6) реализован удачно, а именно: в качестве класса допустимых решений $\mathbf{F}$ определено семейство, «накрывающее» истинную функцию регрессии (11.3), т. е.

$$\mathbf{F} = \{\theta_0 \psi_0(X) + \theta_1 \psi_1(X) + \ldots + \theta_m \psi_m(X)\}_{\Theta \in \Gamma}, \tag{11.6}$$

и, следовательно,

$$f(X) = \mathbf{E}(\eta | \xi = X) \in \mathbf{F}. \tag{11.7}$$

Модель, определяемую соотношениями и условиями (11.1), (11.2), (11.4) и (11.5), будем называть линейным (относительно оцениваемых параметров) нормальным вариантом идеализированной схемы регрессионной зависимости (идеализация, как было отмечено, заключается в постулировании редко выполняющихся в статистической практике допущений (11.7) и (11.2)).

11.1.1. Основные свойства оценок метода наименьших квадратов. Напомним (см. гл. 7—9, а также [14, п. 8.6.3]), что оценки $\widehat{\Theta}$ неизвестных параметров $\Theta = (\theta_0, \theta_1, \ldots, \theta_m)'$, уча-

ствующих в аналитической записи искомой функции регрессии $f(X; \Theta)$, определяются, в соответствии с методом наименьших квадратов, из условия минимизации (по $\widehat{\Theta}$) выборочного критерия адекватности $\widehat{\Delta}_n = \Delta_n(\widehat{\Theta})$, построенного на базе *квадратичной* функции потерь $\rho(u)$ (см. в § 5.2 формулу (5.4′) и п. 1). Применительно к рассматриваемой в данном параграфе схеме это приводит к задаче минимизации (по $\widehat{\Theta}$) выражения:

$$\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = n\Delta_n(\widehat{\Theta}) = \sum_{i=1}^{n}\left(y_i - \sum_{k=0}^{m}\widehat{\theta}_k\,\psi_k(X_i)\right)^2 =$$

$$= (Y - \mathbf{X}\widehat{\Theta})'(Y - \mathbf{X}\widehat{\Theta}) = Y'Y - 2\widehat{\Theta}'\mathbf{X}'Y + \widehat{\Theta}'\mathbf{X}'\mathbf{X}\widehat{\Theta}. \qquad (11.8)$$

При получении правой части (11.8) использовалось что $\widehat{\Theta}'\mathbf{X}'Y = (\widehat{\Theta}'\mathbf{X}'Y)' = Y'\mathbf{X}\widehat{\Theta}$. Дифференцируя (11.8) по $\widehat{\Theta}$ и приравнивая полученный вектор-столбец производных $\partial(\widehat{\varepsilon'\varepsilon})/\partial\widehat{\Theta}$ к вектору $\mathbf{0}$, состоящему из одних нулей, приходим к системе уравнений относительно $\widehat{\Theta} = (\widehat{\theta}_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_m)'$:

$$-2\mathbf{X}'Y + 2\mathbf{X}'\mathbf{X}\widehat{\Theta} = 0$$

или

$$\mathbf{X}'\mathbf{X}\widehat{\Theta} = \mathbf{X}'Y,$$

откуда получаем[1]

$$\widehat{\Theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y. \qquad (11.9)$$

Перед тем, как перейти к описанию основных свойств мнк-оценок $\widehat{\Theta}$, выразим их, подставляя в (11.9) значения Y, представленные в виде (11.4), через истинные значения параметров Θ и регрессионные остатки ε:

$$\widehat{\Theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\Theta + \varepsilon) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\Theta +$$

$$+ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} = \Theta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \qquad (11.10)$$

---

[1] Легко убедиться, что решение (11.9), являющееся, вообще говоря, лишь *стационарной точкой* квадратичной формы (11.8), в данном случае является ее *точкой минимума.* Для этого достаточно воспользоваться тождеством $(Y - \mathbf{X}\Theta)'(Y - \mathbf{X}\Theta) = (Y - \mathbf{X}\widehat{\Theta})'(Y - \mathbf{X}\widehat{\Theta}) + (\widehat{\Theta} - \Theta)'\mathbf{X}'\mathbf{X}(\widehat{\Theta} - \Theta)$, из которого видно, что левая часть достигает минимума при $\Theta = \widehat{\Theta}$.

Используя (11.10), легко получить следующие статистические характеристики для мнк-оценок $\widehat{\Theta}$.

Несмещенность мнк-оценок $\widehat{\Theta}$. Применяя оператор теоретического усреднения к левой и правой частям (11.10), получаем

$$\mathbf{E}\widehat{\Theta} = \Theta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\cdot\mathbf{E}\varepsilon,$$

что, если учесть $\mathbf{E}\varepsilon = \mathbf{0}$, и доказывает несмещенность оценок $\widehat{\Theta}$[1].

Ковариационная матрица мнк-оценок $\widehat{\Theta}$. Как известно, точность оценок, их эффективность [14, п. 8.1.5] определяются характером их выборочного распределения, и, в частности, мерой их *случайного разброса* относительно истинных значений оцениваемых параметров, который мы наблюдали бы при повторениях выборок и принятой процедуры оценивания. В свою очередь эта мера случайного разброса значений оценок $\widehat{\Theta}$ относительно истинных значений Θ определяется в первую очередь их дисперсиями и ковариациями, т. е. их ковариационной матрицей $\Sigma_{\widehat{\Theta}} = \mathbf{E}\,[(\widehat{\Theta} - \Theta)\,(\widehat{\Theta} - \Theta)']$. Подсчитаем ковариационную матрицу $\Sigma_{\widehat{\Theta}}$, используя (11.10) для выражения разности $\widehat{\Theta} - \Theta$:

$$\Sigma_{\widehat{\Theta}} = \mathbf{E}\,[(\widehat{\Theta}-\Theta)\,(\widehat{\Theta}-\Theta)'] = \mathbf{E}\,\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon]\,[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon]'\} =$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\cdot\mathbf{E}\,(\varepsilon\varepsilon')\cdot\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

(при переходе к правой части мы воспользовались правилом транспонирования произведения матриц и симметричностью матрицы $(\mathbf{X}'\mathbf{X})^{-1}$. Если теперь учесть, что ковариационная матрица $\mathbf{V} = \mathbf{E}\,(\varepsilon\varepsilon')$ регрессионных остатков пропорциональна единичной (см. (11.5)), то в конечном счете получим

$$\Sigma_{\widehat{\Theta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{11.11}$$

где $\sigma^2 = \mathbf{D}\varepsilon$, а $\mathbf{X}$ — определенная выше матрица плана

[1]Здесь и в дальнейшем мы пользуемся возможностью выносить выражения, зависящие от матрицы плана $\mathbf{X}$, из-под знака теоретического усреднения $\mathbf{E}$ в соответствии с известным правилом: $\mathbf{E}\,(c\xi) = c\cdot\mathbf{E}\xi$ (где c — неслучайная величина), поскольку величины $\psi_k\,(X_i)$ в рассматриваемой идеализированной схеме не являются случайными.

Оценка $\widehat{\sigma^2}$ для дисперсии σ^2 регрессионных остатков. Оценка для σ^2, полученная с помощью метода максимального правдоподобия [14, п. 8.6.1], имеет вид [119, формула (4.8)]

$$\widehat{\sigma^2_{\text{мп}}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{m} \widehat{\theta}_k \psi_k (X_i) \right)^2 = \frac{1}{n} (Y - \mathbf{X}\widehat{\Theta})' \cdot (Y - \mathbf{X}\Theta),$$

однако она оказывается *смещенной*. В частности, можно показать [119, § 3.3], что, взяв в качестве оценки для σ^2 величину

$$\widehat{\sigma^2} = \frac{n}{n-m-1} \widehat{\sigma^2_{\text{мп}}} = \frac{1}{n-m-1} (y - \mathbf{X}\widehat{\Theta})' \cdot (Y - \mathbf{X}\widehat{\Theta}), \qquad (11.12)$$

мы добъемся *несмещенного* оценивания этого параметра.

Состоятельность оценок $\widehat{\Theta}$ и $\widehat{\sigma^2}$. Она определяется структурой матрицы плана $\mathbf{X}$. Пожалуй, наиболее удобным (для приложений) условием состоятельности оценок $\widehat{\Theta}$ и $\widehat{\sigma^2}$ является следующее [119, § 3.2]: оценки $\widehat{\Theta}$ и $\widehat{\sigma^2}$ состоятельны тогда и только тогда, когда наименьшее собственное значение матрицы $\mathbf{X}'\mathbf{X}$ стремится к бесконечности при $n \to \infty$.

Оптимальность оценок $\widehat{\Theta}$ и $\widehat{\sigma^2}$. Можно показать, что в условиях рассматриваемой идеализированной регрессионной схемы оценки $\widehat{\theta}_k$ ($k=0, 1, 2, \ldots, m$) и $\widehat{\sigma^2}$, определяемые соотношениями (11.9) и (11.12), являются эффективными [119, § 3.2], т. е. *имеют минимальную дисперсию среди всех несмещенных оценок*. Тем же свойством обладает и величина $\widehat{f}(X) = f(X; \widehat{\Theta}) = \widehat{\theta}_0\psi_0(X) + \widehat{\theta}_1\psi_1(X) + \ldots + \widehat{\theta}_m \psi_m(X)$, рассматриваемая как оценка истинной функции регрессии $f(X; \Theta) = \theta_0\psi_0(X) + \theta_1\psi_1(X) + \ldots + \theta_m \psi_m(X)$.

Распределение оценок регрессионных параметров. Характер случайного варьирования оценок $\widehat{\Theta}$, $\widehat{\sigma^2}$ и $\widehat{f}(X)$ около оцениваемых ими величин соответственно Θ, σ^2 и $f(X)$ описывается *лишь приближенно* их ковариационной матрицей и дисперсиями. Исчерпывающую же информацию о характере этого случайного варьирования доставляют соответствующие законы распределения вероятностей.

Нетрудно убедиться, что в рамках рассматриваемой в данном параграфе идеализированной схемы справедливы следующие утверждения:

а) оценки $\hat{\Theta}$ подчиняются $(m+1)$-мерному нормальному[1] распределению $N_{m+1}(\Theta, \Sigma_{\hat{\Theta}})$ с вектором средних значений Θ и с ковариационной матрицей (11.11), т. е.

$$\hat{\Theta} \in N_{m+1}(\Theta; \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}); \qquad (11.13)$$

б) случайная величина $(\hat{\Theta} - \Theta)'\mathbf{X}'\mathbf{X}(\hat{\Theta} - \Theta)/\sigma^2$ подчиняется χ^2-распределению с $m+1$ степенями свободы, т. е.

$$\frac{1}{\sigma^2}(\hat{\Theta} - \Theta)' \cdot \mathbf{X}'\mathbf{X} \cdot (\hat{\Theta} - \Theta) \in \chi^2(m+1); \qquad (11.14)$$

в) оценки $\hat{\Theta}$ и $\hat{\sigma}^2$ являются *статистически независимыми*;

г) случайная величина $(n-m-1)\hat{\sigma}^2/\sigma^2$ подчиняется χ^2-распределению с $n-m-1$ степенями свободы, т. е.

$$\frac{n-m-1}{\sigma^2}\hat{\sigma}^2 \in \chi^2(n-m-1). \qquad (11.15)$$

Действительно, поскольку $\hat{\Theta} = CY$ (см. (11.9)), а Y в силу (11.2′) и (11.4) подчиняется n-мерному нормальному распределению, то утверждение (11.13) следует непосредственно из того, что линейные комбинации нормально распределенных величин также распределены нормально [20, теорема 2.4.1]. Утверждение (11.14) является прямым следствием (11.11) и теоремы 3.3.3 из [20]. Статистическая независимость оценок $\hat{\Theta}$ и $\hat{\sigma}^2$ и утверждение (11.15) следуют, например, из теоремы 8.2.2 [20]. Полное доказательство сформулированных результатов можно найти также в [119, § 3.4].

З а м е ч а н и е. Обращаем внимание читателя на тот факт, что допущение (11.2)—(11.2′) о нормальном характере распределения регрессионных остатков ε используется *лишь при выводе распределений оценок* (11.9), т. е. при получении результатов (11.13)—(11.15). Остальные свойства рассматриваемых оценок: несмещенность, состоятельность, оптимальность (но только в классе *линейных* несмещенных оценок), вид ковариационной матрицы (11.11) — остаются в силе и при отказе от

---

[1]Для простоты полагаем здесь, что матрица плана $\mathbf{X}$ имеет *полный ранг* (т. е. ранг $(\mathbf{X}) = m+1$). Если это не так (т. е. если ранг $(\mathbf{X}) < m+1$), то все сформулированные в данном параграфе результаты остаются в силе с поправкой на соответствующее снижение числа степеней свободы или размерности того пространства, о котором идет речь. Более подробно об этом см. [119, § 3.8].

нормальности остатков ε (достаточно потребовать их одинаковой распределенности, независимости и существования конечных дисперсий $D\varepsilon_i = \sigma^2 < \infty$).

11.1.2. Решение основных задач по оценке точности регрессионной модели. В § В.6 сформулированы три основные задачи анализа точности регрессионной модели. Эти задачи сводятся к умению указать такие гарантированные (с заданной доверительной вероятностью P) предельные величины погрешностей, за пределы которых мы не выйдем, если вместо неизвестных истинных значений параметров θ_k ($k = 0, 1, \ldots, m$), функции регрессии $f(X)$ (при заданном значении предиктора X) и анализируемого результирующего показателя $\eta(X) = (\eta|\xi = X)$ (тоже при заданном значении предиктора X) будем использовать *их оценки* соответственно $\widehat{\theta}_k$, $\widehat{f}(X) = f(X; \widehat{\Theta})$ и снова $\widehat{f}(X) = f(X; \widehat{\Theta})$.

Описанные в п. 11.1.1 свойства оценок $\widehat{\Theta}$ позволяют предложить следующий способ конструирования этих предельных гарантированных величин погрешностей.

Погрешность $\delta_{P,n}(\theta_k)$ в оценивании параметра θ_k. Воспользуемся нормальной распределенностью оценки $\widehat{\theta}_k$ (см. (11.13)) и знанием ее среднего значения $\mathbf{E}\,\widehat{\theta}_k = \theta_k$, (см. свойство несмещенности оценок $\widehat{\Theta}$ в п. 11.1.1) и дисперсии $\widehat{\mathbf{D}\theta_k} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}_{kk}$ (см. (11.11); здесь $(\mathbf{X}'\mathbf{X})^{-1}_{kk}$ обозначает k-й диагональный элемент матрицы $(\mathbf{X}'\mathbf{X})^{-1}$). Это, с учетом статистической независимости $\widehat{\Theta}$ и $\widehat{\sigma^2}$ и (11.15), позволяет утверждать, что величина

$$t_1(n-m-1) = (\widehat{\theta}_k - \theta_k)/\widehat{\sigma}\,[(\mathbf{X}'\mathbf{X})^{-1}_{kk}]^{1/2}$$

подчиняется t-распределению, или распределению Стьюдента [14, п. 6.2.2], с $n-m-1$ степенями свободы. Следовательно, если задана величина доверительной вероятности P, то, отыскав по табл. П.6 $100(1-P)/2\%$-ную точку $t_{\frac{1-P}{2}}(n-m-1)$, мы можем с вероятностью P гарантировать выполнение неравенства

$$\frac{|\widehat{\theta}_k - \theta_k|}{\widehat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}} < t_{\frac{1-P}{2}}(n-m-1),$$

или, что то же,

$$\left.\begin{array}{l} |\hat{\theta}_k - \theta_k| < \delta_{P,n}(\theta_k) \\ \text{где} \quad \delta_{P,n}(\theta_k) = t_{\frac{1-p}{2}}(n-m-1)\,\hat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{kk}}. \end{array}\right\} \qquad (11.16)$$

Погрешность $\delta_{P,n}(y_{ср}(X))$ в оценивании функции регрессии (при заданном значении X предиктора). Обозначим $\Psi(X) = (\psi_0(X), \psi_1(X), \ldots, \psi_m(X))'$ и вычислим, учитывая (11.11) и $\mathbf{E}\hat{f}(X) = f(X)$ (что следует из несмещенности оценок $\hat{\theta}_k$, $k = 0, 1, \ldots, m$), дисперсию оценки $\hat{f}(X)$:

$$\mathbf{D}\hat{f}(X) = \mathbf{E}(\hat{f}(X) - f(X))^2 = \mathbf{E}\,[\Psi'(X)\cdot(\hat{\Theta} - \Theta)]^2 =$$
$$= \Psi'(X)\cdot\mathbf{E}\,[(\hat{\Theta} - \Theta)\cdot(\hat{\Theta} - \Theta)']\cdot\Psi(X) =$$
$$= \sigma^2\,[\Psi'(X)\cdot(\mathbf{X}'\mathbf{X})^{-1}\cdot\Psi(X)]. \qquad (11.17)$$

Учитывая полученное выражение для $\mathbf{D}\hat{f}(X)$, несмещенность и нормальную распределенность оценки $\hat{f}(X) = \Psi'(X)\cdot\hat{\Theta}$ (как линейной функции от нормально распределенных случайных величин $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_m$) и ее статистическую независимость от $\hat{\sigma}^2$ (см. п. 11.1.1), а также (11.15), получаем факт $t(n-m-1)$-распределенности случайной величины

$$\hat{t}_2(n-m-1) = \frac{\hat{f}(X) - f(X)}{\hat{\sigma}\sqrt{\Psi'(X)\cdot(\mathbf{X}'\mathbf{X})^{-1}\Psi(X)}}.$$

Следовательно, можно гарантировать с заданной величиной доверительной вероятности P выполнение неравенства

$$|\hat{t}_2(n-m-1)| < t_{\frac{1-P}{2}}(n-m-1), \qquad (11.18)$$

где $t_\alpha(\nu)$ — $100\alpha\%$-ная точка распределения Стьюдента с ν степенями свободы (определяется из табл. П.6). Очевидно, что (11.18) равносильно неравенству

$$\left.\begin{array}{l} |\hat{f}(X) - f(X)| < \delta_{P,n}(y_{ср}(X)), \\ \text{где} \\ \delta_{P,n}(y_{ср}(X)) = \\ = t_{\frac{1-P}{2}}(n-m-1)\cdot\hat{\sigma}\cdot\sqrt{\Psi'(X)\cdot(\mathbf{X}'\mathbf{X})^{-1}\cdot\Psi(X)}. \end{array}\right\} \qquad (11.18')$$

выполнение которого гарантируется с заданной доверительной вероятностью P.

Погрешность $\delta_{P.n}(\eta(X))$ в восстановлении «индивидуального» значения результирующего показателя (при заданном значении предиктора). В данном случае нас интересует оценка сверху для величины погрешности, которую мы можем допустить, восстанавливая (прогнозируя) с помощью $\hat{f}(X)$ неизвестное значение результирующего показателя $\eta(X)$ при заданном фиксированном значении X предикторной переменной. Другими словами, статистическому анализу следует подвергнуть величину $\hat{f}(X)-\eta(X)$. Принимая во внимание $\mathbf{E}\hat{f}(X) = \mathbf{E}\eta(X) = f(X)$, (11.2) и статистическую независимость величин $\varepsilon(X) = \eta(X) - f(X)$ и $\hat{\varepsilon} = (y_1 - \hat{f}(X_1),\ y_1 - \hat{f}(X_2), \ldots, y_n - \hat{f}(X_n))'$,
имеем

$$\mathbf{E}(\hat{f}(X)-\eta(X))^2 = \mathbf{E}[(\hat{f}(X)-f(X))-(\eta(X)-f(X))]^2 =$$

$$= \mathbf{E}(\hat{f}(X)-f(X))^2+\mathbf{E}\varepsilon^2(X) = \mathbf{D}\hat{f}(X)+\sigma^2,$$

где дисперсия $\mathbf{D}\hat{f}(X)$ определяется соотношением (11.17).

Учитывая $(0, \sigma^2\cdot[1+\Psi'(X)(\mathbf{X}'\mathbf{X})^{-1}\Psi(X)])$-нормальную распределенность погрешности $\hat{f}(X)-\eta(X)$ и ее статистическую независимость от $\hat{\sigma}^2$ (см. в) в п. 11.1.1), получаем факт $t(n-m-1)$-распределенности случайной величины

$$\hat{t}_3(n-m-1) = (\hat{f}(X)-\eta(X))/\hat{\sigma}\cdot\sqrt{1+\Psi'(X)\cdot(\mathbf{X}'\mathbf{X})^{-1}\Psi(X)}.$$

Следовательно, задавшись величиной доверительной вероятности P и определив из табл. П.6 величину $100(1-P)/2\%$-ной точки распределения Стьюдента с $n-m-1$ степенями свободы $t_{\frac{1-P}{2}}(n-m-1)$, можно гарантировать (с вероятностью P) выполнение неравенства

$$|\hat{t}_3(n-m-1)| < t_{\frac{1-P}{2}}(n-m-1),$$

или, что то же,

$$|\hat{f}(X)-\eta(X)| < t_{\frac{1-P}{2}}(n-m-1)\cdot\hat{\sigma}\times$$

$$\times\sqrt{1+\Psi'(X)\cdot(\mathbf{X}'\mathbf{X})^{-1}\Psi(X)}. \qquad (11.19)$$

Замечание (о построении доверительных областей). Хотя и весьма редко, но возникают ситуации (в частности, при статистическом анализе и управлении ходом технологического процесса), когда требуется дать оценки погрешностей вида (11.18′) и (11.19), которые были бы справедливыми с заданной доверительной вероятностью P *одновременно для целого множества* A значений предикторной переменной X, т. е. для всех $X \in$ A. В таких ситуациях говорят о построении *доверительных областей* для $f(X)$ (или $\eta(X)$) при $X \in$ A. Обобщение оценок (11.18′), (11.19) на этот случай можно найти в [119, гл. 5].

11.1.3. Случаи линейной (по предикторным переменным) и полиномиальной регрессии. Воспользуемся полученными в предыдущем пункте рекомендациями для анализа точности моделей линейной и полиномиальной регрессии. Нас будет интересовать, в частности, конкретизация формул (11.18′) и (11.19) в этих случаях.

Па́рная линейная регрессия. Рассматривается модель вида (11.1), в которой размерность предиктора $p = 1$, а система базисных функций задается соотношениями: $\psi_0(x) \equiv 1$; $\psi_1(x) = x$, так что в конечном счете анализируется зависимость вида $\eta = \theta_0 + \theta_1\xi + \varepsilon$ или

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i. \tag{11.4'}$$

Тогда

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_n \end{pmatrix}; \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^{n} x_i^2 \end{pmatrix};$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix};$$

$$\mathbf{X}'Y = \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix},$$

где $\bar{x} = \sum_{i=1}^{n} x_i/n$ и $\bar{y} = \sum_{i=1}^{n} y_i/n$ — арифметические средние значения наблюденных величин предиктора x и результирующего показателя y соответственно.

Далее

$$\Psi'(X)(\mathbf{X}'\mathbf{X})^{-1}\Psi(X) = (1, x)\cdot\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\times$$

$$\times\begin{pmatrix}\frac{1}{n}\sum_{i=1}^{n}x_i^2 & -\bar{x}\\ -\bar{x} & 1\end{pmatrix}\begin{pmatrix}1\\ x\end{pmatrix} = \frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

так что (в соответствии с (11.18′) и (11.19)) с вероятностью P будем иметь выполнение неравенств (при заданном фиксированном значении предиктора x):

$$(\widehat{\theta}_0+\widehat{\theta}_1 x)-(\theta_0+\theta_1 x)\,|< t_{\frac{1-P}{2}}(n-2)\times$$

$$\times\widehat{\sigma}\cdot\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}; \qquad (11.18'')$$

$$(\widehat{\theta}_0+\widehat{\theta}_1 x)-\eta(x)\,|< t_{\frac{1-P}{2}}(n-2)\times$$

$$\times\widehat{\sigma}\cdot\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}, \qquad (11.19')$$

где $t_\alpha(\nu)$, как и прежде, $100\alpha\%$-ная точка распределения Стьюдента с ν степенями свободы; $\widehat{\sigma^2}$ — оценка остаточной дисперсии (см. (11.12)), т. е.

$$\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\widehat{\theta}_0-\widehat{\theta}_1 x_i)^2,$$

а $\widehat{\theta}_0$ и $\widehat{\theta}_1$ — оценки наименьших квадратов неизвестных коэффициентов θ_0 и θ_1 (см. (11.9)), т. е.

$$\widehat{\theta}_0=\overline{y}-\widehat{\theta}_1\cdot\overline{x};$$

$$\widehat{\theta}_1=\frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

Из (11.18″) и (11.19′), в частности, видно, что: а) величина погрешности и в том и в другом случае зависит от того, при каком именно значении предиктора x производится оценка, причем эта погрешность (и соответственно ширина доверительного интервала) увеличивается по мере удаления заданного значения x от среднего арифметического $\overline{x}$ и наблюденных значений предикторной переменной; б) погрешность оценивания неизвестной функции регрессии $f(x)=\theta_0+\theta_1 x$ пропорциональна величине $\widehat{\sigma}/\sqrt{n}$ и, следовательно, неограниченно убывает с ростом объема выборки (n), по которой производится оценивание; в) погрешность оценивания *индивидуального* (*а не среднего*) значения результирующего показателя $\eta(x)=(\eta|\xi=x)$ с помощью вычисленной по методу наименьших квадратов функции регрессии $\widehat{f}(x)=\widehat{\theta}_0+\widehat{\theta}_1 x$ при неограниченном увеличении объема выборки (т. е. при $n\to\infty$) в отличие от предыдущей погрешности не убывает до нуля, но стремится, как это и должно быть (в соответствии с допущением (11.2)), к величине $(1-P)/2\%$-ной точки $(0,\sigma^2)$-нормального распределения, т. е. к величине $u_{\frac{1+P}{2}}\cdot\sigma$ (поскольку, как известно [71], $t(\nu)$-распределение сходится к стандартному нормальному при $\nu\to\infty$, а следовательно, при $n\to\infty$, $t_{\frac{1-P}{2}}(n-2)\to u_{\frac{1+P}{2}}$, где u_α — q-квантиль стандартного нормального распределения; а $\widehat{\sigma^2}\to\sigma^2=D\varepsilon$ в силу состоятельности $\widehat{\sigma^2}$).

Множественная линейная регрессия. Обобщим модель (11.4′) на случай p ($p>1$) предикторных переменных $X=(x^{(1)}, x^{(2)}, \ldots, x^{(p)})'$. Запишем исследуемую модель $\eta=\theta_0+\theta_1\xi^{(1)}+\ldots+\theta_p\xi^{(p)}+\varepsilon$ в терминах *центрированных* наблюденных переменных, т. е.

$$y_i-\overline{y}=\theta_1\cdot(x_{(i)}^{(1)}-\overline{x}^{(1)})+\ldots+\theta_p(x_i^{(p)}-\overline{x}^{(p)})+\varepsilon_i. \quad (11.4'')$$

Тогда получаем следующую интерпретацию общих обозначений п. 11.1.1 и 11.1.2:

$$Y = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \cdots \\ y_n - \bar{y} \end{pmatrix}; \ \psi_k(X) = x^{(k+1)} - \bar{x}^{(k+1)} \quad (k = 0, 1, \ldots, p-1);$$

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} - \bar{x}^{(1)} & x_1^{(2)} - \bar{x}^{(2)} & \ldots & x_1^{(p)} - \bar{x}^{(p)} \\ x_2^{(1)} - \bar{x}^{(1)} & x_2^{(2)} - \bar{x}^{(2)} & \ldots & x_2^{(p)} - \bar{x}^{(p)} \\ \cdots & \cdots & \cdots & \cdots \\ x_n^{(1)} - \bar{x}^{(1)} & x_n^{(2)} - \bar{x}^{(2)} & \ldots & x_n^{(p)} - \bar{x}^{(p)} \end{pmatrix};$$

$$\mathbf{X}'\,\mathbf{X} = (n\hat{\sigma}_{kj})_{k,\, j=\overline{1,\, p}}, \text{ где } \hat{\sigma}_{kj} = \frac{1}{n} \sum_{i=1}^{n} \left(x_i^{(k)} - \bar{x}^{(k)}\right)\left(x_i^{(j)} - \bar{x}^{(j)}\right) -$$

выборочные ковариации переменных $x^{(k)}$ и $x^{(j)}$ [14, п. 5.6.7]. Соответственно

$$(\mathbf{X}'\,\mathbf{X})^{-1} = \frac{1}{n}\,(\widehat{\sigma^{(k\cdot j)}})_{k,\, j=\overline{1,\, p}},$$

где $\widehat{\sigma^{(k\cdot j)}}$ — (k, j)-й элемент матрицы, обратной к выборочной ковариационной матрице предиктора X.

Таким образом,

$$\Psi'(X)\cdot(\mathbf{X}'\,\mathbf{X})^{-1}\,\Psi(X) = \frac{1}{n}\cdot\left(x^{(1)} - \bar{x}^{(1)} \ x^{(2)} - \bar{x}^{(2)} \ \ldots \ x^{(p)} - \right.$$

$$\left. - \bar{x}^{(p)}\right)\cdot \begin{pmatrix} \widehat{\sigma^{(1.1)}} & \widehat{\sigma^{(1.2)}} & \ldots & \widehat{\sigma^{(1.p)}} \\ \widehat{\sigma^{(2.1)}} & \widehat{\sigma^{(2.2)}} & \ldots & \widehat{\sigma^{(2.p)}} \\ \cdots & \cdots & \cdots & \cdots \\ \widehat{\sigma^{(p.1)}} & \widehat{\sigma^{(p.2)}} & \ldots & \widehat{\sigma^{(p\cdot p)}} \end{pmatrix} \begin{pmatrix} x^{(1)} - \bar{x}^{(1)} \\ x^{(2)} - \bar{x}^{(2)} \\ \cdots \\ x^{(p)} - \bar{x}^{(p)} \end{pmatrix} =$$

$$= \frac{1}{n} \sum_{k=1}^{p} \sum_{j=1}^{p} \left(x^{(k)} - \bar{x}^{(k)}\right)\left(x^{(j)} - \bar{x}^{(j)}\right)\cdot \widehat{\sigma^{(k\cdot j)}}.$$

Поэтому в соответствии с (11.18′) и (11.19) с вероятностью P можно гарантировать выполнение следующих неравенств

при заданном фиксированном векторном значении $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})'$ предикторных переменных:

$$\left|\sum_{k=1}^{p} \widehat{\theta}_k (x^{(k)} - \bar{x}^{(k)}) - \sum_{k=1}^{p} \theta_k (x^{(k)} - \bar{x}^{(k)})\right| <$$

$$< t_{\frac{1-P}{2}}(n-p) \cdot \frac{\widehat{\sigma}}{\sqrt{n}} \cdot \sqrt{\sum_{k=1}^{p} \sum_{j=1}^{p} (x^{(k)} - \bar{x}^{(k)})(x^{(j)} - \bar{x}^{(j)}) \sigma^{(k.j)}}\,; \qquad (11.18''')$$

$$\left|\sum_{k=1}^{p} \widehat{\theta}_k (x^{(k)} - \bar{x}^{(k)}) - (\eta(X) - \mathbf{E}\,\eta)\right| < t_{\frac{1-P}{2}}(n-p) \cdot \widehat{\sigma} \times$$

$$\times \sqrt{1 + \frac{1}{n} \sum_{k=1}^{p} \sum_{j=1}^{p} (x^{(k)} - \bar{x}^{(k)})(x^{(j)} - \bar{x}^{(j)}) \widehat{\sigma}^{(k.j)}}\,. \qquad (11.19'')$$

Предварительный переход к центрированным переменным обусловил то, что правые части (11.18‴) и (11.19″) *не учитывают погрешности в оценке неизвестного теоретического среднего* $a_\eta = \mathbf{E}_\eta$. Модификация левых и правых частей (11.18‴) и (11.19″), соответствующая возвращению к исходной (нецентрированной) записи модели, заключается в следующем: в левые части неравенств надо прибавить (внутри прямых скобок) величину $\bar{y} - a_\eta$, а в правые части (под знак радикала) — соответственно единицу и $1/n$, изменив в обоих случаях число степеней свободы у процентной точки t-распределения на $n - p - 1$. Нетрудно увидеть, что записанные в таком модифицированном виде неравенства (11.18‴) и (11.19″) дают в качестве своего частного случая при $p = 1$ неравенства (11.18″) и (11.19′).

Полиномиальная регрессия. Рассмотрим случай *скалярного* (т. е. одномерного, $p = 1$) предиктора x, и пусть искомая функция регрессии $f(x)$ принадлежит классу алгебраических полиномов степени m, т. е.

$$f(x) = b_0 + b_1 x + \ldots + b_m x^m.$$

Ограничимся для определенности случаем $m = 2$ (переход к общему случаю $m > 2$ осуществляется очевидным образом без каких-либо затруднений) и представим функцию регрессии в системе базисных функций $\{\psi_0(x), \psi_1(x), \psi_2(x)\}$, являющихся *ортогональными* (на совокупности наблюденных

значений предиктора $(x_1, x_2, \ldots, x_n)$ *полиномами Чебышева* (см. гл. 7, а также [10, с. 131; 77, с. 275]), т. е.

$$\eta = \theta_0 \psi_0(\xi) + \theta_1 \psi_1(\xi) + \theta_2 \psi_2(\xi) + \varepsilon,$$

где

$$\psi_0(x) \equiv 1;$$

$$\psi_1(x) = x - \bar{x};$$

$$\psi_2(x) = x^2 - \frac{\sum_{i=1}^{n} x_i^3 - \bar{x} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} (x - \bar{x}) - \frac{1}{n} \sum_{i=1}^{n} x_i^2.$$

Взаимная ортогональность полиномов $\psi_j(x)$ (на системе наблюдений $x_1, x_2, \ldots, x_n$) означает, что

$$\sum_{=1}^{n} \psi_j(x_i) \cdot \psi_k(x_i) = 0 \text{ при } j \neq k,$$

что обеспечивает диагональность матрицы $\mathbf{X}'\mathbf{X}$. В частности, имеем

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum_{i=1}^{n} \psi_0^2(x_i) & 0 & 0 \\ 0 & \sum_{i=1}^{n} \psi_1^2(x_i) & 0 \\ 0 & 0 & \sum_{i=1}^{n} \psi_2^2(x_i) \end{pmatrix};$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \left(\sum_{i=1}^{n} \psi_0^2(x_i)\right)^{-1} & 0 & 0 \\ 0 & \left(\sum_{i=1}^{n} \psi_1^2(x_i)\right)^{-1} & 0 \\ 0 & 0 & \left(\sum_{i=1}^{n} \psi_2^2(x_i)\right)^{-1} \end{bmatrix}$$

Полученные в соответствии с (11.9) мнк-оценки $\widehat{\theta}_k = \sum_{i=1}^{n} y_i \cdot \psi_k(x_i) / \sum_{i=1}^{n} \psi_k^2(x_i)$ $(k = 0, 1, 2)$ оказываются статисти-

чески взаимно независимыми и имеют дисперсии $\mathbf{D}\widehat{\theta}_k = \sigma^2 \cdot (\sum_{i=1}^{n} \psi_k^2(x_i))^{-1}$.

Учитывая, что в данном случае

$$\Psi'(x) \cdot (\mathbf{X}'\mathbf{X})^{-1} \cdot \Psi(x) = \sum_{k=0}^{2} \left(\frac{\psi_k^2(x)}{\sum_{i=1}^{n} \psi_k^2(x_i)} \right),$$

можно (в соответствии с (11.18′) и (11.19)) гарантировать с вероятностью P выполнение неравенств (при заданном значении x):

$$\left| \sum_{x=0}^{2} \widehat{\theta}_k \cdot \psi_k(x) - f(x) \right| < t_{\frac{1-p}{2}}(n-3) \cdot \widehat{\sigma} \times$$

$$\times \sqrt{\sum_{k=0}^{2} \frac{\psi_k^2(x)}{\sum_{i=1}^{n} \psi_k^2(x_i)}};$$

$$\left| \sum_{k=0}^{2} \widehat{\theta}_k \cdot \psi_k(x) - \eta(x) \right| < t_{\frac{1-p}{2}}(n-3) \cdot \widehat{\sigma} \times$$

$$\times \sqrt{1 + \sum_{k=0}^{2} \frac{\psi_k^2(x)}{\sum_{i=1}^{n} \psi_k^2(x_i)}},$$

где

$$\widehat{\sigma}^2 = \frac{1}{n-3} \sum_{i=1}^{n} (y_i - \widehat{\theta}_0 - \widehat{\theta}_1 \psi_1(x_i) - \widehat{\theta}_2 \psi_2(x_i))^2,$$

а $t_\alpha(\nu)$ — 100α%-ная точка распределения Стьюдента с ν степенями свободы (определяется из табл. П.6).

11.2. Нелинейный нормальный вариант идеализированной схемы регрессионной зависимости

В данном параграфе исследуются вопросы точности регрессионного анализа применительно к общей параметрической модели регрессии, в которой наблюденные значения y_i и $X_i = (x_i^{(1)}, ..., x_i^{(p)})'$ соответственно результирующего показателя η и объясняющей (неслучайной) переменной $\xi = (\xi^{(1)}, ..., \xi^{(p)})'$ связаны соотношениями

$$y_i = f(X_i;\ \theta_0,\ \theta_1, ..., \theta_m) + \varepsilon_i,\ i = 1, 2, ..., n. \qquad (11.20)$$

При этом постулируется выполнение следующих допущений:

1) выбранный исследователем класс допустимых решений **F** содержит в себе искомую функцию регрессии $f(X; \Theta) = \mathbf{E}(\eta | \xi = X)$, т. е.

$$f(X; \Theta) \in \mathbf{F}; \qquad (11.21)$$

2) регрессионные остатки $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ несмещены относительно нуля, взаимно некоррелированы и одинаково $(0; \sigma^2)$-нормально распределены, т. е.

$$\boldsymbol{\varepsilon} \in N(0; \sigma^2 \cdot \mathbf{I}_n); \qquad (11.22)$$

3) искомая функция регрессии $f(X; \Theta)$ *нелинейно* зависит от подлежащих статистическому оцениванию параметров $\Theta = (\theta_0, \theta_1, ..., \theta_m)$, однако характер этой зависимости *достаточно гладкий* (например, существуют всевозможные вторые производные от $f(X; \Theta)$ по параметрам θ_k, θ_j, $k, j = 0, 1, ..., m$).

Напомним, что главным (*принципиальным*) пунктом «идеализации» анализируемой схемы является первый, т. е. допущение (11.21). Два других носят, скорее, полутехнический характер.

Относительная сложность решения различных вопросов точности регрессионного анализа (по сравнению с предыдущим линейным вариантом) состоит в том, что в данном случае мнк-оценки $\widehat{\Theta}$ неизвестных параметров Θ определяются не в виде явных аналитических выражений (ср. с (11.9)), а лишь в ходе итерационных алгоритмических процедур (см. гл. 9), что существенно затрудняет исследование их свойств. В основе обычно используемых в данной схеме подходов — разложение в ряд Тейлора (по параметрам Θ в окрестности наилучшей

оценки $\widehat{\Theta}$) оптимизируемого критерия адекватности $\Delta(f(X;\Theta)) = \Delta(\Theta)$ (см. гл. 5 и 9) и искомой функции регрессии $f(X;\Theta)$.

11.2.1. Основные свойства мнк-оценок. Поскольку мнк-оценки $\widehat{\Theta}$ параметра Θ в условиях (11.20)—(11.22) совпадают с оценками максимального правдоподобия [14, п. 8.6.1; 25, гл. 4], то мы можем воспользоваться общими результатами о свойствах последних. Из них, в частности, следует, что мнк-оценки $\widehat{\Theta}$ регрессионных коэффициентов Θ модели (11.20) являются (при условии соблюдения упомянутых условий) *состоятельными*, *асимптотически-несмещенными*, *асимптотически-эффективными* и *асимптотически-нормальными* (асимптотика по $n \to \infty$).

Для того чтобы перейти непосредственно к решению задач анализа точности нелинейной регрессионной модели, нам необходимо получить предварительно выражение, аналогичное (11.11), для ковариационной матрицы $\Sigma_{\widehat{\Theta}}$ мнк-оценок $\widehat{\Theta}$.

С этой целью воспользуемся разложением функции регрессии $f(X_i;\Theta)$ в ряд Тейлора в окрестности точки $\Theta = \widehat{\Theta}$, (где $\widehat{\Theta}$ — мнк-оценка параметра Θ, полученная с помощью процедур, описанных в гл. 9), ограничиваясь линейными членами разложения:

$$f(X_i;\Theta) \approx f(X_i;\widehat{\Theta}) + (\Theta - \widehat{\Theta})' \cdot \left.\frac{\partial f(X_i;\Theta)}{\partial \Theta}\right|_{\Theta=\widehat{\Theta}}$$

или, в обозначениях гл. 9 (см. п. 9.3.1):

$$f_i(\Theta) \approx f_i(\widehat{\Theta}) + (\Theta - \widehat{\Theta})' \cdot \dot{f}_i, \tag{11.23}$$

где $f_i(\Theta) = f(X_i;\Theta)$ и $\dot{f}_i = (\partial f_i/\partial\theta_0, \ldots, \partial f_i/\partial\theta_m)'$.

Выражение (11.23) для произвольного значения X дает:

$$f(X;\Theta) - f(X;\widehat{\Theta}) \approx (\Theta - \widehat{\Theta})' \cdot \dot{f}. \tag{11.24}$$

Введение обозначений $\tilde{\psi}_k(X) = \left.\frac{\partial f(X;\Theta)}{\partial \theta_k}\right|_{\Theta=\widehat{\Theta}}$,

$$\tilde{\Theta} = \Theta - \widehat{\Theta};\ \tilde{f}(X;\tilde{\Theta}) = f(X;\tilde{\Theta}+\widehat{\Theta}) - f(X;\widehat{\Theta});$$

$$\tilde{\Psi}(X) = \dot{f} = (\tilde{\psi}_0(X), \tilde{\psi}_1(X), \ldots, \tilde{\psi}_m(X))'$$

позволяет записать выражение (11.24) в виде

$$\tilde{f}(X;\tilde{\Theta}) \approx \tilde{\Theta}' \cdot \tilde{\Psi}(X) \tag{11.24'}$$

и свести, таким образом, нелинейную модель (11.20) к ее аппроксимации линейной схемой, рассмотренной в предыдущем параграфе.

Решение задач анализа точности регрессионной модели основано на исследовании точности оценок $\widehat{\Theta}$ и отклонений $f(X; \Theta) - f(X; \widehat{\Theta})$. Примем во внимание приближенные соотношения (11.24) и (11.24′), равенство ковариационных матриц оценок $\widehat{\Theta}$ и $\widetilde{\Theta}$ (т. е. $\Sigma_{\hat{\Theta}} = \Sigma_{\tilde{\Theta}}$) и возможность использования обычного приближенного приема вычисления вторых моментов статистических оценок, когда в полученные выражения для этих моментов, зависящие от неизвестных значений оцениваемых параметров, вставляются *их оценки*. Тогда можно, опираясь на результаты предыдущего параграфа (см. также [25, § 7.5]), получить следующие выражения для ковариационной матрицы $\Sigma_{\widehat{\Theta}}$ оценок $\widehat{\Theta}$ и для ее оценки $\widehat{\Sigma}_{\widehat{\Theta}}$:

$$\Sigma_{\widehat{\Theta}} \approx \sigma^2 \cdot \mathbf{M}_{\Theta}^{-1}; \quad \widehat{\Sigma}_{\widehat{\Theta}} \approx \widehat{\sigma^2} \cdot \mathbf{M}_{\widehat{\Theta}}^{-1}, \qquad (11.25), (11.26)$$

где

$$\widehat{\sigma^2} = \frac{1}{n-m-1} \sum_{i=1}^{n} (y_i - f(X_i; \widehat{\Theta}))^2; \qquad (11.27)$$

$$\mathbf{M}_{\Theta^*} = \sum_{i=1}^{n} \mathbf{M}_{\Theta^*}(X_i);$$

$$\mathbf{M}_{\Theta^*}(X_i) = \begin{bmatrix} \left(\frac{\partial f(X_i; \Theta)}{\partial \theta_0}\right)^2 & \frac{\partial f(X_i; \Theta)}{\partial \theta_0} \cdot \frac{\partial f(X_i; \Theta)}{\partial \theta_1} & \cdots & \frac{\partial f(X_i; \Theta)}{\partial \theta_0} \cdot \frac{\partial f(X_i, \Theta)}{\partial \theta_m} \\ \frac{\partial f(X_i; \Theta)}{\partial \theta_1} \cdot \frac{\partial f(X_i; \Theta)}{\partial \theta_0} & \left(\frac{\partial f(X_i; \Theta)}{\partial \theta_1}\right)^2 & \cdots & \frac{\partial f(X_i; \Theta)}{\partial \theta_1} \cdot \frac{\partial f(X_i; \Theta)}{\partial \theta_m} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f(X_i; \Theta)}{\partial \theta_m} \cdot \frac{\partial f(X_i; \Theta)}{\partial \theta_0} & \frac{\partial f(X_i; \Theta)}{\partial \theta_m} \cdot \frac{\partial f(X_i; \Theta)}{\partial \theta_1} & \cdots & \left(\frac{\partial f(X_i; \Theta)}{\partial \theta_m}\right)^2 \end{bmatrix},$$

причем производные, участвующие в выражении элементов матрицы $\mathbf{M}_{\Theta^*}(X_i)$, берутся в точке $\Theta = \Theta^*$, т. е.

$$\frac{\partial f(X_i;\ \Theta)}{\partial \theta_k} \cdot \frac{\partial f(X_i;\ \Theta)}{\partial \theta_j} = \frac{\partial f(X_i;\ \Theta)}{\partial \theta_k}\bigg|_{\Theta=\Theta^*} \cdot \frac{\partial f(X_i,\ \Theta)}{\partial \theta_j}\bigg|_{\Theta=\Theta^*}$$

$k, j = 0, 1, \ldots, m;\ \Theta^* = \Theta, \widehat{\Theta}.$

11.2.2. Решение основных задач по оценке точности нелинейной регрессионной модели. Подчеркнем два главных отличия данного случая от линейного, рассмотренного в § 11.1. Во-первых, используемые для построения доверительных интервалов свойства состоятельных мнк-оценок $\widehat{\Theta}$ — несмещенность, оптимальность, нормальность, а также свойства б), в) и г) из п. 11.1.1 справедливы лишь в *асимптотическом* (*по* $n \to \infty$) *смысле*. Во-вторых, следует учитывать *приближенный характер* базовых соотношений (11.24) и соответственно (11.25) и (11.26). Следует признать, что возможны различные уточнения описываемого здесь приближенного подхода [161]. Однако вряд ли они существенно усовершенствуют предлагаемые в данном параграфе практические рекомендации: ведь даже так называемые точные критерии и доверительные интервалы *на практике оказываются всего лишь приближенными* (они точны лишь в той мере, в какой соблюдаются в реальной ситуации те идеализированные допущения, на которых строятся соответствующие статистические выводы). Поэтому, говоря о том, что интересующая нас погрешность не превзойдет определенной величины с доверительной вероятностью, например, равной 0,95, мы должны всегда отдавать себе отчет в приближенном характере подобных заключений.

Учитывая сделанное замечание, читатель может использовать для приближенного решения трех основных задач оценки точности нелинейной регрессионной модели соотношения (11.16), (11.18′) и (11.19) предыдущего параграфа с заменой: матрицы $(\mathbf{X}'\mathbf{X})$ матрицей $\mathbf{M}_{\widehat{\Theta}}$ (см. (11.28)); оценки $\widehat{\sigma}^2$ оценкой, подсчитанной по формуле (11.27); вектор-столбца $\Psi(X)$ вектор-столбцом

$$\widetilde{\Psi}(X) = \boldsymbol{f} = \left(\frac{\partial f(X;\ \widehat{\Theta})}{\partial \widehat{\theta}_0},\ \frac{\partial f(X;\ \widehat{\Theta})}{\partial \widehat{\theta}_1},\ \ldots,\ \frac{\partial f(X;\ \widehat{\Theta})}{\partial \widehat{\theta}_m} \right).$$

11.3. Исследование точности регрессионной модели в реалистической ситуации

Неточный выбор общего вида функции регрессии, приводящий к нарушению базового допущения (11.21), на которое существенно опираются все выводы по оцениванию точности регрессионной модели, может заключаться как в неполном или избыточном представлении набора объясняющих переменных $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, так и в искажении *самой структуры модели*. Наиболее неприятные последствия влечет второй тип ошибки[1]. В этом можно убедиться при рассмотрении примера 6.2, а также примера, представленного в табл. 6.2 и на рис. 6 2. Действительно, анализируя данные табл. 6.1 (в которой представлены результаты расчетов по примеру 6.2), мы видим, в частности, что при использовании *формально-аппроксимационных* вариантов регрессионной модели (т. е. в ситуации $f(X) \notin \mathbf{F}$) оценки среднеквадратической ошибки остатков $(\widehat{\sigma})$, полученные по формуле (11.27) по данным *той же самой выборки, по которой вычислены и оценки* $\widehat{\Theta}$ неизвестных параметров модели, дают более чем в 3 раза заниженные (по сравнению с действительными) значения (см. графы 4 и 6). Более того, из примера, представленного на рис. 6.2 (и в табл. 6.2), следует, что значение выборочного критерия адекватности $\widehat{\Delta}_n$ (пропорционального величине $\widehat{\sigma^2}$) вообще *может быть нулевым* (!), в то время как ошибки восстановления неизвестных значений функции регрессии f или результирующего показателя η по заданной величине предиктора x могут быть практически сколь угодно велики (ср. поведение $\widehat{f_a}(x)$ с $f(x)$ и $\eta(x)$ при $x \in [7; 14]$ и при $x > 17$).

Подмеченные в рассмотренных примерах особенности *аппроксимационных вариантов регрессионных моделей* (так мы будем называть варианты, в которых истинная функция регрессии $f(X) \notin \mathbf{F} = \{f_a(X; \Theta)\}_{\Theta \in \Gamma}$) приводят к следующим основным положениям исследования точности статистических выводов в регрессионном анализе в данной ситуации:

---

[1]Влияние неполного или избыточного представления набора объясняющих переменных на свойства оценок и соответственно на точность статистических выводов в регрессионном анализе (при правильном определении структуры модели) может быть учтено в рамках строгих математических конструкций (см., например, [119, гл. 6]).

1) при анализе точности аппроксимационных вариантов регрессионных моделей не следует претендовать на построение сколько-нибудь точных доверительных интервалов ни для неизвестных значений параметров Θ (они, как правило, не имеют в данной ситуации самостоятельной содержательной интерпретации), ни для функции регрессии $f(X)$ или результирующего показателя $\eta(X)$ (поскольку, пользуясь аппроксимацией $\widehat{f_a}(X)$, отличающейся по структуре от истинной функции регрессии $f(X)$, мы не можем иметь достоверной априорной информации о вероятностной природе остатков $\widehat{\varepsilon_i} = y_i - \widehat{f_a}(X_i)$);

2) имеющуюся выборку наблюдений $\widetilde{B}_n = \{(X_1, y_1), \ldots, (X_n, y_n)\}$ целесообразно разбить (одним или несколькими различными способами) на две непересекающиеся подвыборки объемов n_1 и n_2 $(n_1 + n_2 = n)$: *обучающую* $\widetilde{B}_n^{\text{об}}$, на основании наблюдений которой строятся мнк-оценки $\widehat{\Theta}^{(n_1)}$ неизвестных параметров аппроксимационной функции регрессии $f_a(X; \Theta)$, и *экзаменующую* (или *контрольную*) $\widetilde{B}_{n_2}^{\text{экз}}$, по наблюдениям которой оцениваются основные характеристики точности анализируемой модели (в первую очередь регрессионные остатки $\widehat{\varepsilon_i} = y_i - f_a(X_i; \widehat{\Theta}^{(n_1)})$);

3) основной (и, по существу, единственной) характеристикой точности аппроксимационного варианта регрессионной модели является оценка $\widehat{\sigma}$ среднеквадратической ошибки аппроксимации σ, вычисляемая по формуле

$$\widehat{\sigma}^2 = \frac{1}{\sum\limits_{j=1}^{k} n_{2j} - m - 1} \sum_{j=1}^{k} \sum_{(X_i\, y_i) \in \widetilde{B}_{n_{2j}}^{\text{экз}}} (y_i - f(X_i; \widehat{\Theta}^{(n_{1j})}))^2, \quad (11.27')$$

где подразумевается, что имеющаяся выборка наблюдений $\widetilde{B}_n = \{(X_1, y_1), \ldots, (X_n, y_n)\}$ разбита k различными способами на две непересекающиеся подвыборки — обучающую $\widetilde{B}_{n_{1j}}^{\text{об}}$ и экзаменующую (или контрольную) $\widetilde{B}_{n_{2j}}^{\text{экз}}$ соответственно объемов n_{1j} и n_{2j} $(j = 1, 2, \ldots, k)$, а мнк-оценки $\widehat{\Theta}^{(n_{1j})}$ неизвестных параметров Θ построены только по n_{1j} данным, входящим в состав обучающей выборки $\widetilde{B}_{n_{1j}}^{\text{об}}$. Знание $\widehat{\sigma}$ позволяет оценить максимально возможную погрешность аппроксимации неизвестной функции регрессии $f(X)$ (в пределах обследо-

ванного диапазона значений X) приблизительно величиной порядка $\pm 2\widehat{\sigma}/\sqrt{n}$, а результирующего показателя $\eta(X)$ — величиной порядка $\pm 2\widehat{\sigma}$;

4) следует проявлять известную сдержанность и осторожность при использовании аппроксимационных вариантов регрессионных моделей для решения задач интерполяции и (особенно) экстраполяции, т. е. при восстановлении неизвестного значения функции регрессии $f(X)$ или результирующего показателя $\eta(X)$ по значению предиктора X, *лежащему вне статистически обследованной области значений объясняющих переменных* (см. также гл. 6 и 8).

Поясним подробнее конструктивную реализацию положений 2) и 3) на примере использования широко применяемого метода *скользящего экзамена*[1]. Определим n разбиений исходной выборки $\widetilde{B}_n = \{(X_1, y_1), \ldots, (X_n, y_n)\}$ на обучающую ($B^{\text{об}}_{n_{1j}}$) и экзаменующую ($B^{\text{экз}}_{n_{2j}}$) следующим образом:

$$\begin{cases} \widetilde{B}^{\text{об}}_{n_{1j}} = \{(X_1, y_1), \ldots, (X_{j-1}, y_{j-1}), (X_{j+1}, y_{j+1}), \ldots, (X_n, y_n)\}; \\ B^{\text{экз}}_{n_{2j}} = \{(X_j, y_j)\}, \; j = 1, 2, \ldots, n. \end{cases}$$

Таким образом: $n_{1j} = n - 1$ и $n_{2j} = 1$ для всех $j = 1, 2, \ldots, n$; j-й вариант обучающей выборки содержит все наблюдения исходной выборки $\widetilde{B}_n$ кроме одного — (X_j, y_j); соответственно j-й вариант экзаменующей выборки содержит единственное наблюдение — (X_j, y_j). Применение к такой последовательности обучающих и экзаменующих выборок формулы (11.27′) дает:

$$\widehat{\sigma}^2 = \frac{1}{n-m-1} \sum_{i=1}^{n} (y_i - f(X_i, \widehat{\Theta}^{(n_{1i})}))^2. \qquad (11.27'')$$

Величина среднеквадратической погрешности $\widehat{\sigma}$, подсчитанная с помощью метода скользящего экзамена (11.27″), в аппроксимационных схемах регрессии оказывается, как правило, существенно больше аналогичной характеристики, вычисленной с помощью обычной формулы (11.27).

З а м е ч а н и е (по поводу вычислительной реализации метода скользящего экзамена). На первый взгляд реализация

[1]Этот метод (в зарубежной литературе он называется «методом складного ножа», или «jackknife method») является одним из вариантов реализации общего подхода, впервые предложенного, по-видимому, в связи с задачей устранения смещения в статистических оценках (см. [65]).

метода скользящего экзамена связана с многократным повторением громоздких вычислений на ЭВМ. Действительно, процедура предусматривает n-кратное вычисление оценок $\widehat{\Theta}^{(n_{11})}$, $\widehat{\Theta}^{(n_{12})}$, ..., $\widehat{\Theta}^{(n_{1n})}$, n^2-кратное вычисление выборочных функций регрессии $f_a(X_i; \widehat{\Theta}^{(n_{1j})})$ $(i, j = 1, 2, ..., n)$ и т. д. Однако непосредственный анализ основных формул метода наименьших квадратов в случае линейного вида аппроксимирующих функций $f_a(X; \Theta) = \theta_0\psi_0(X) + ... + \theta_m\psi_m(X) = \Theta'\psi(X)$ (см. формулы (11.3), (11.9)—(11.12)) позволяет установить полезные соотношения между интересующими нас характеристиками, подсчитанными *по всей выборке* $\widetilde{B}_n$, и теми же характеристиками, подсчитанными по выборке, в которой нет наблюдения (X_i, y_i):

$$\widehat{\Theta}^{(n_{1i})} = \widehat{\Theta} + \frac{y_i - \widehat{\Theta}'\Psi(X_i)}{1-q_i} \cdot (\mathbf{X}'\mathbf{X})^{-1} \cdot \Psi(\mathbf{X}_i), \tag{11.28}$$

где $q_i = \Psi'(X_i) \cdot (\mathbf{X}'\mathbf{X})^{-1} \cdot \Psi(X_i)$*;

$$\widehat{\Theta}^{(n_{1i})'}\Psi(X_i) = \widehat{\Theta}'\Psi(X_i) + q_i \frac{y_i - \widehat{\Theta}'\Psi(X_i)}{1-q_i}; \tag{11.29}$$

$$\widehat{\varepsilon}_i(\widehat{\Theta}^{(n_{1i})}) = \frac{\widehat{\varepsilon}_i(\widehat{\Theta})}{1-q_i}, \text{ где } \widehat{\varepsilon}_i(\Theta) = y_i - f(X_i; \Theta); \tag{11.30}$$

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i^2(\widehat{\Theta}^{(n_{1i})}) = \sum_{i=1}^{n} \frac{1}{(1-q_i)^2} \widehat{\varepsilon}_i^2(\Theta); \tag{11.31}$$

$$(n-1)\widehat{\Delta}_{n-1}(\widehat{\Theta}^{(n_{1i})}) = n\widehat{\Delta}_n(\widehat{\Theta}) - \frac{1+q_i}{1-q_i}(y_i - \widehat{\Theta}'\Psi(X_i))^2. \tag{11.32}$$

Соотношения (11.28)—(11.31) позволяют избежать многократной вычислительной «прогонки» процедур метода наименьших квадратов на различных вариантах обучающей выборки за счет пересчета значений $\widehat{\Theta}^{(n_{1i})}$, $f(X_i; \widehat{\Theta}^{(n_{1i})}) = \widehat{\Theta}^{(n_{1i})'}\Psi(X_i)$ и т. д. по соответствующим характеристикам, подсчитанным по наблюдениям *всей* выборки $\widetilde{B}_n$.

---

*При самых естественных ограничениях на структуру матрицы плана $\mathbf{X}$ величина q_i, несмотря на зависимость от X_i, ведет себя при $n \to \infty$ (при *фиксированной* размерности $m+1$ оцениваемого параметра) как n^{-1}; если же и $m \to \infty$, то $q_i \sim (m+1)/n$.

ВЫВОДЫ

1. Завершив построение оценки (или аппроксимации) $f(X; \widehat{\Theta}) = \widehat{f}(X)$ для неизвестной истинной функции регрессии $f(X) = \mathbf{E}(\eta|\xi = X)$, исследователь должен по возможности определить ту гарантированную (с заданной доверительной вероятностью P) величину погрешности, за пределы которой он не выйдет, восстанавливая неизвестные значения параметров θ_j, истинной функции регрессии $f(X)$ или анализируемого результирующего показателя $\eta(X) = (\eta|\xi = X)$ по значениям их оценок соответственно $\widehat{\theta}_j$, $\widehat{f}(X)$ и $\widehat{f}(X)$.

2. Решающим моментом во всей процедуре исследования точности статистических выводов в регрессионном анализе является соотношение между истинной функцией регрессии $f(X)$ и выбранным исследователем параметрическим классом допустимых решений $\mathbf{F} = \{f_a(X; \Theta)\}_{\Theta \in \Gamma}$. Если класс $\mathbf{F}$ выбран удачно (т. е. если $f(X) \in \mathbf{F}$), то исследователь находится в рамках *идеализированной схемы* и при некоторых дополнительных априорных сведениях о природе регрессионных остатков $\varepsilon(X) = \eta - f(X)$ имеет возможность дать достаточно точный ответ на все три основных вопроса анализа точности регрессионной модели (см. § 11.1, 11.2).

Если исследователь не может гарантировать включения $f(X) \in \mathbf{F}$ (что и бывает в большинстве практических ситуаций), то он находится в рамках *реалистической схемы* и может, в лучшем случае, дать лишь весьма грубую оценку $\widehat{\sigma}$ для среднеквадратической погрешности аппроксимации.

3. Решение задач оценки точности *линейной* (по оцениваемым параметрам) модели регрессии в рамках идеализированной схемы опирается на такие свойства мнк-оценок $\widehat{\Theta}$ неизвестных параметров регрессии Θ, как *состоятельность*, *несмещенность*, *оптимальность* и *нормальность*, а также на умение вычислить (в терминах матрицы наблюдений, или матрицы плана $\mathbf{X}$) ковариационную матрицу оценок $\widehat{\Theta}$.

4. Решение задачи оценки точности *нелинейной* модели регрессии в рамках идеализированной схемы опирается на те же свойства мнк-оценок $\widehat{\Theta}$, справедливые в данном случае, правда, лишь в асимптотическом (по $n \to \infty$) смысле, а также на разложение функции регрессии $f(X; \Theta)$ в ряд Тейлора (по параметру Θ) в окрестности точки $\Theta = \widehat{\Theta}$ (где $\widehat{\Theta}$ — мнк-оценка параметра Θ) и на умение вычислить (в терминах частных

производных $\partial f(X;\Theta)/\partial\theta_j$) ковариационную матрицу оценок $\widehat{\Theta}$.

5. Анализ точности регрессионной модели в рамках *реалистической* схемы сводится к вычислению оценки $\widehat{\sigma}$ для среднеквадратической погрешности аппроксимации σ по обычной формуле выборочной остаточной дисперсии. Однако остатки $\widehat{\varepsilon}_i = y_i - f(X_i;\widehat{\Theta})$, на основании которых подсчитывается величина $\widehat{\sigma}^2$ (см. формулы (11.27), (11.27′) и (11.27″)), следует вычислять лишь для тех наблюдений (X_i, y_i), которые не вошли в состав данных, по которым рассчитываются мнк-оценки $\widehat{\Theta}$. Следовательно, анализ точности регрессионной модели в реалистической ситуации предусматривает необходимость предварительного разбиения имеющихся исходных статистических данных на две непересекающиеся выборки — *обучающую* (по данным которой строятся мнк-оценки $\widehat{\Theta}$) и *экзаменующую* (по данным которой оцениваются регрессионные остатки $\widehat{\varepsilon}_i = y_i - f(X_i;\widehat{\Theta})$).

6. Частным (и весьма распространенным) вариантом оценивания $\widehat{\sigma}^2$ по экзаменующей выборке является метод *скользящего экзамена*, в котором в качестве n экзаменующих выборок последовательно используется каждое из n исходных наблюдений (X_1,y_1), ..., $(X_n y_n)$, а соответственно остальные $n-1$ наблюдений используются в качестве обучающих выборок. При этом оценка $\widehat{\sigma}^2$ вычисляется по формуле (11.27″).

Глава 12. СТАТИСТИЧЕСКИЙ АНАЛИЗ АВТОРЕГРЕССИОННЫХ ДИНАМИЧЕСКИХ ЗАВИСИМОСТЕЙ

В данной главе рассматривается случай, когда исследуется поведение *единственной* случайной переменной x *во времени*. Исходной статистической базой для такого исследования является ряд значений

$$x_1, x_2, \ldots, x_n \tag{12.1}$$

исследуемой переменной, зарегистрированных в последовательные моменты времени соответственно $t_1, t_2, \ldots, t_n$.

Последовательность наблюдений типа (12.1) принято называть *временным рядом*. Он имеет два главных отличия от рассматриваемых наблюдений анализируемого признака, образующих *случайные выборки*: а) образующие временной ряд наблюдения $x_1, x_2, \ldots, x_n$, рассматриваемые как случайные величины, *не являются взаимно независимыми*, и, в частности, значение, которое мы получим в момент времени t_k ($k = 1, 2, \ldots, n$), может существенно зависеть от того, какие значения были зарегистрированы до этого момента времени; б) наблюдения временного ряда (в отличие от элементов случайной выборки), вообще говоря, *не образуют стационарной последовательности*, т. е. закон распределения вероятностей k-го члена временного ряда (случайной величины $x_k = x(t_k)$) не остается одним и тем же при изменении его номера k; в частности, от t_k могут зависеть основные числовые характеристики случайной переменной x_k — ее среднее значение $\mathbf{E}x(t_k)$ и дисперсия $\mathbf{D}x(t_k)$ (функцию от аргумента t, описывающую зависимость $\mathbf{E}x(t)$ от времени, часто называют *трендом* временного ряда).

Статистическое исследование последовательностей вида (12.1) осуществляется с помощью специального раздела математической статистики — *анализа временных рядов*. В данной главе рассматриваются модели лишь одного частного типа — *модели авторегрессии*. Базовая идея, на которой эти модели строятся, как раз и заключается в использовании вышеуказанной особенности (а) временных рядов, и, в частности, в постулировании возможности восстановления значения анализируемой переменной x в момент времени t (т. е. величины $x(t)$) по ее же собственным значениям, зафиксированным в предыдущие моменты времени $t - 1, t - 2, \ldots$ (отсюда и происхождение названия моделей).

Более полное и основательное освещение аппарата анализа временных рядов приведено в [21, 28, 41, 66, 80, 144].

12.1. Дискретные динамические модели

Одним из наиболее наглядных, простых и в то же время весьма часто и плодотворно используемых динамических моделей являются модели вида

$$x_t = \sum_{\nu=1}^{m} \theta_\nu x_{t-\nu} + \sum_{j=0}^{p} \gamma_j \varepsilon_{t-j}, \qquad (12.2)$$

где θ_ν и γ_j — параметры, подлежащие оцениванию по наблюдениям (12.1), а ε_t — случайные величины, удовлетворяющие условиям

$$\mathbf{E}(\varepsilon_t)=0,\ \ \mathbf{E}(\varepsilon_t\varepsilon_{t'}) = \begin{cases}\sigma^2, \text{ если } t=t';\\ 0, \text{ если } t\neq t.\end{cases} \tag{12.2'}$$

Если $p=0$, то (12.2) называют *моделью авторегрессии*. Если же в правой части (12.2) равно нулю первое слагаемое, то говорят о модели *скользящего среднего*. При $p>0$ и $m>0$ соотношения вида (12.2) называют *смешанной моделью авторегрессии и скользящего среднего*.

В отличие от регрессионных моделей, где случайность чаще всего описывает погрешность наблюдения, модель (12.2)—(12.2′) предполагает, что уже сам исследуемый объект описывается вероятностной моделью, т. е. что анализируемая переменная x — случайная величина. Типичной областью применения модели (12.2)—(12.2′) являются эконометрические исследования (см. гл. 14). При учете трендов оказывается удобным рассматривать вместо (12.2) ее несколько усложненный вариант:

$$x_t=\sum_{\nu=1}^{m_1}\theta_{\nu 1}x_{t-\nu}+\sum_{\nu=1}^{m_2}\theta_{\nu 2}\psi_\nu(t)+\sum_{j=0}^{p}\gamma_j\varepsilon_{t-j}, \tag{12.3}$$

где $\psi_\nu(t)$ — *заданные* функции. Нередко «точные» значения x_t оказываются недоступными. Тогда модели (12.2) или (12.3) приходится дополнять предположениями, что для анализа доступны лишь «зашумленные» наблюдения

$$y_t=x_t+\tilde{\varepsilon}_t. \tag{12.4}$$

Обычно предполагается, что

$$\mathbf{E}(\tilde{\varepsilon}_t)=0,\ \mathbf{E}(\tilde{\varepsilon}_t^2)=\tilde{\sigma}^2,\ \mathbf{E}(\tilde{\varepsilon}_t\tilde{\varepsilon}_{t'})=0 \text{ при } t\neq t'; \tag{12.4}$$

$$\mathbf{E}(\varepsilon_t\tilde{\varepsilon}_{t'})=0.$$

Как будет ясно из дальнейшего, численные аспекты задач оценивания, порождаемых моделями (12.2)—(12.3), не вызывают каких-либо затруднений, так как опираются на стандартный аппарат метода наименьших квадратов (см. гл. 7—9). Однако изучение статистических свойств соответствующих оценок приводит к целому ряду довольно сложных проблем. С большинством из них приходится сталкиваться уже при изучении простейшего варианта модели (12.2) — авторегрессии первого порядка. Именно поэтому авторегрессия первого порядка и будет достаточно подробно изучена в следующем параграфе.

12.2. Авторегрессия первого порядка

12.2.1. Нормально распределенные «возмущения». Рассмотрим простейший вариант модели (12.2):

$$x_t = \theta x_{t-1} + \varepsilon_t, \quad t = 1, 2, \dots, n. \tag{12.5}$$

Предположим вначале, что случайные величины ε_t («возмущения») распределены по нормальному закону и $\mathbf{E}(\varepsilon_t) = 0$, $\mathbf{E}(\varepsilon_t^2) = \sigma^2$, $\mathbf{E}(\varepsilon_t \varepsilon_{t'}) = 0$ при $t \neq t'$. Пусть также задано начальное условие x_0.

Соотношение (12.5) удобно представить в виде

$$\mathbf{L}X = \boldsymbol{\varepsilon}, \tag{12.6}$$

где $X = (x_1, \dots, x_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1 + \theta x_0, \varepsilon_2, \dots, \varepsilon_n)'$, а матрица $\mathbf{L}$ имеет размерность $n \times n$ и представляется в виде

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\theta & 1 & 0 & \dots & 0 \\ 0 & -\theta & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}. \tag{12.7}$$

Из (12.6) следует, что вектор X имеет нормальное распределение со средним $\overline{X}$, удовлетворяющим уравнению

$$\mathbf{L}\overline{X} = X_0, \tag{12.8}$$

где $X_0 = (x_0, 0, \dots, 0)'$. Нетрудно видеть, что $\mathbf{E}x_t = x_0\theta^t$. Ковариационная матрица $\mathbf{D}(X)$ вектора X равна:

$$\mathbf{D}(X) = \sigma^2 \mathbf{L}\mathbf{L}'. \tag{12.9}$$

Заметим, что $|\mathbf{L}| = |\mathbf{L}\mathbf{L}'| = 1$. Из (12.8) и (12.9) следует, что распределение вектора наблюдений X описывается плотностью

$$p(X|\sigma^2, \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(x_t - \theta x_{t-1})^2\right].$$

Как обычно, для получения оценок максимального правдоподобия $\widehat{\sigma_n^2}$ и $\widehat{\theta_n}$ необходимо максимизировать функцию $p(X/\sigma^2, \theta)$ по σ^2 и θ. Нетрудно видеть, что решениями этой экстремальной задачи являются

$$\widehat{\theta_n} = M_n^{-1} y_n, \tag{12.10}$$

где

$$\mathcal{M}_n = \sum_{t=1}^{n} x_{t-1}^2; \quad y_n = \sum_{t=1}^{n} x_t x_{t-1};$$

$$\widehat{\sigma_n^2} = n^{-1} \sum_{t=1}^{n} (x_t - \widehat{\theta_n} x_{t-1})^2. \qquad (12.11)$$

Таким образом, с вычислительной точки зрения мы имеем дело с простейшей линейной задачей метода наименьших квадратов. Изучение же статистических свойств оценок (12.10) и (12.11) оказывается заметно сложнее, чем для классической линейной регрессии.

12.2.2. Асимптотические свойства оценок. Изучение предельного поведения (при $n \to \infty$) оценки $\widehat{\theta_n}$ опирается на анализ поведения последовательностей $\sum_{t=1}^{n} x_{t-1}^2$ и $\sum_{t=1}^{n} \varepsilon_t x_{t-1}$, так как

$$\widehat{\theta_n} - \theta = \sum_{t=1}^{n} \varepsilon_t x_{t-1} \Bigg/ \sum_{t=1}^{n} x_{t-1}^2.$$

Оказывается, что *при любом* θ *оценка* $\widehat{\theta_n}$ *является состоятельной* (*т. е. сходится по вероятности* к θ).

Предельное распределение оценки $\widehat{\theta_n}$ оказывается различным при $|\theta| < 1$, $|\theta| > 1$ и $|\theta| = 1$.

Если $|\theta| < 1$, то говорят об *устойчивом случае*. При этом $\lim\limits_{n\to\infty} \mathbf{E} x_n = 0$.

В устойчивом случае случайная величина $\sqrt{n}\,(\widehat{\theta_n} - \theta)$ *имеет в пределе нормальное распределение с нулевым средним и дисперсией* $\mathbf{D} = (\sum_{s=0}^{\infty} \theta^{2s})^{-1} = 1 - \theta^2$. Обратим внимание, что при $|\theta| < 1$ величина $\sum_{t=1}^{n} x_t^2/n$ сходится по вероятности к $\sigma^2(1-\theta^2)^{-1}$.

Если $|\theta| > 1$, то мы имеем дело с неустойчивым («взрывным») случаем. При этом $\lim\limits_{n\to\infty} \mathbf{E} x_n = \infty$.

В неустойчивом случае случайная величина $|\theta|^n (\widehat{\theta_n} - \theta) \times \times (\theta^2 - 1)^{-1}$ *имеет в пределе распределение Коши* [14, п. 6.1.10] *с параметрами* (*0*, *1*).

Наиболее сложным оказывается промежуточный случай, когда $|\theta| = 1$. Доказано, что при $|\theta| = 1$ *случайная величина* $n(\widehat{\theta}_n - \theta)/\sqrt{2}$ *имеет в пределе распределение с плотностью*:

$$p(u) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \frac{\partial \varphi}{\partial x}(t, -tx)\, \mathrm{d}\, t,$$

где

$$\varphi(t, z) = \mathrm{e}^{\sqrt{2}\,\theta i t}\left(\cos 2(iz)^{1/2} - \frac{\theta i t}{(2iz)^{1/2}} \sin 2(iz)^{1/2}\right)^{-1/2}$$

12.2.3. Произвольное распределение «возмущений». Оценка (12.10) формально совпадает с мнк-оценкой. Так же как и в случае регрессионных задач, кажется естественным использовать ее и при распределениях «возмущений» ε_t, не совпадающих с нормальным, а лишь удовлетворяющих условиям (12.2′), дополненным требованием их одинаковой распределенности и независимости. Оказывается [21, 155], что в этом случае *оценка (12.10) является состоятельной при любых* θ, *причем*:

1) *при* $|\theta| < 1$ *случайная величина* $\sqrt{n}(\widehat{\theta}_n - \theta)$ *асимптотически нормально распределена с нулевым средним и дисперсией* $1 - \theta^2$;

2) *при* $|\theta| > 1$ *случайная величина* $|\theta|^n(\widehat{\theta}_n - \theta)(\theta - 1)^{-2}$ *имеет некоторое предельное распределение, характеристики которого зависят от закона распределения «возмущений»*;

3) *аналогичное утверждение имеет место и для* $|\theta| = 1$, *но для случайной величины* $n(\widehat{\theta}_n - \theta)/\sqrt{2}$. Во всех рассмотренных выше случаях асимптотическая «точность» (точнее, предельное распределение выписанных выражений) *не зависит от дисперсии «возмущений»*.

Подводя итог результатам, изложенным в данном и предыдущем пунктах, можно сказать, что при $|\theta| < 1$ скорость сходимости порядка $n^{-\frac{1}{2}}$; при $|\theta| > 1$ — порядка $|\theta|^{-n}(1 - \theta)^2$ и при $|\theta| = 1$ — порядка $n^{-1}\sqrt{2}$. Таким образом, устойчивая авторегрессия первого порядка с точки зрения оценивания является «наихудшей».

Для всех рассмотренных выше случаев (как в п. 12.2.2, так и в п. 12.2.3) оценка (12.11) является состоятельной, причем при достаточно больших n предельное распределение случайной величины $\widehat{\sigma}_n^2/\sigma^2$ близко к χ^2-распределению с n степенями свободы.

12.3. Авторегрессия произвольного порядка

В полной аналогии с (12.6) для авторегрессии m-го порядка можно записать, что

$$\mathbf{L}X = \boldsymbol{\varepsilon}, \tag{12.12}$$

где $X = (x_1, \dots, x_n)'$, $\boldsymbol{\varepsilon} = \left(\varepsilon_1 + \sum_{\nu=1}^{m} \theta_\nu x_{1-\nu},\ \varepsilon_2 + \sum_{\nu=2}^{m} \theta_\nu x_{2-\nu}, \dots, \varepsilon_n\right)'$, $x_0, x_{-1}, \dots, x_{1-m}$ — начальные условия,

$$\mathbf{L} = \mathbf{I}_n - \sum_{\nu=1}^{m} \theta_\nu \mathbf{L}_1^\nu, \tag{12.13}$$

где

$$\mathbf{L}_1 = \begin{bmatrix} 0 & 0 & 0 & . & . & . & 0 & 0 \\ 1 & 0 & 0 & . & . & . & 0 & 0 \\ 0 & 1 & 0 & . & . & . & 0 & 0 \\ . & & & & . & . & . & . \\ 0 & 0 & 0 & . & . & . & 1 & 0 \end{bmatrix}$$

Вектор X при нормальных и независимых «возмущениях» (см. комментарии к (12.5)) распределен по нормальному закону со средним $\overline{X}$, удовлетворяющим уравнению

$$\mathbf{L}\overline{X} = X_0, \tag{12.14}$$

где $X_0 = \left(\sum_{\alpha=1}^{m} \theta_\alpha x_{1-\alpha}, \prod_{\alpha=2}^{m} \theta_\alpha x_{1-\alpha}, \dots, \theta_m x_{1-m}, 0, \dots, 0\right)'$,

и с ковариационной матрицей

$$\mathbf{D}(X) = \sigma^2 \mathbf{L}\mathbf{L}'. \tag{12.15}$$

Решение уравнения (12.14) (точнее, его k-ю компоненту) можно представить в виде

$$\overline{x}_k = \sum_{j=1}^{r} \sum_{t=0}^{q_j - 1} b_{jt} k^t z_j^k, \tag{12.16}$$

где z_j — корень уравнения

$$z^m = \sum_{\nu=1}^{m} \theta_\nu z^{m-\nu}, \tag{12.17}$$

q_j — кратность этого корня, r — количество различных корней, константы b_{jt} определяются из начальных условий. Уравнение (12.17) является характеристическим для разностного уравнения

$$x_m = \sum_{\nu=1}^{m} \theta_\nu x_{m-\nu},$$

которое, впрочем, эквивалентно матричному уравнению $\mathbf{L}X = 0$. При $m = 1$ решение (12.16), конечно же, совпадает с уже упоминавшимся в предыдущем параграфе решением $\bar{x}_k = x_0\theta^k$.

Из (12.12), (12.14), (12.15) следует, что плотность распределения для вектора X может быть представлена в виде

$$p(X|\sigma^2;\Theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{n}\left(x_t - \sum_{\nu=1}^{m}\theta_\nu x_{t-\nu}\right)^2\right]. \tag{12.18}$$

Так же как и при $m = 1$, полезно иметь в виду, что $|\mathbf{L}| = |\mathbf{LL}'| = 1$.

С учетом (12.18) оценки максимального правдоподобия $\widehat{\Theta}_n = (\widehat{\theta}_{1n}, \ldots, \widehat{\theta}_{mn})'$ и $\widehat{\sigma_n^2}$ имеют вид (ср. с (12.10) и (12.11)):

$$\widehat{\Theta}_n = \mathbf{M}_n^{-1} Y_n, \tag{12.19}$$

где компоненты матрицы $\mathbf{M}_n$ ($m_{\alpha\beta}$) и вектора Y_n (y_α) определяются соотношениями:

$$m_{\alpha\beta} = \sum_{t=1}^{n} x_{t-\alpha} x_{t-\beta}, \quad y_\alpha = \sum_{t=1}^{n} x_t x_{t-\alpha} \quad (\alpha, \beta = \overline{1,m}); \tag{12.20}$$

$$\widehat{\sigma_n^2} = n^{-1}\sum_{t=1}^{n}\left(x_t - \sum_{\alpha=1}^{m}\widehat{\theta}_\alpha x_{t-\alpha}\right)^2.$$

Формулы (12.19) и (12.20) полностью аналогичны (12.10), (12.11), так что для отыскания $\widehat{\theta}_n$ и $\widehat{\sigma_n^2}$ могут быть использо-

ваны стандартные программы метода наименьших квадратов. Следует, однако, иметь в виду, что матрица **M** в задачах авторегрессии, как правило, оказывается плохо обусловленной (см. гл. 8).

Асимптотическое поведение компонент $\widehat{\theta}_{\alpha n}$ векторной оценки $\widehat{\Theta}_n$ определяется значениями корней уравнения (12.17). Наиболее хорошо изучен случай, когда все корни по модулю меньше единицы: $|z_\gamma| < 1$, $\gamma = \overline{1, r}$ (см., например, [21, гл. 5]). При этом предположение о нормальности распределения «возмущений» оказывается несущественным.

Если все корни характеристического уравнения (12.17) по модулю меньше единицы, а случайные «возмущения» ε_t *независимы,* $\mathbf{E}(\varepsilon_t) = 0$, $\mathbf{E}(\varepsilon_t^2) = \sigma^2$, $t = \overline{1, n}$ *и имеют одинаковые распределения, то оценки (12.19) и (12.20) состоятельные.* Более того [21], *случайная величина* $\sqrt{n}(\widehat{\Theta}_n - \Theta)$ *имеет в пределе нормальное распределение с нулевым средним и ковариационной матрицей* $\mathbf{A}^{-1}$, где

$$\mathbf{A} = \sum_{s=0}^{\infty} \mathbf{B}^s \mathbf{I}_1^* \mathbf{B}'^s,$$

$$\mathbf{I}_1^* = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \end{pmatrix}.$$

Так же как и в предыдущем параграфе, асимптотическая дисперсия оценок не зависит от дисперсии «возмущений» ε_t, а целиком определяется истинными значениями параметров Θ.

В практических задачах полезно иметь в виду, что матрица $n^{-1}\sigma^{-2}\mathbf{M}_n$ (см. (12.19)) является состоятельной оценкой матрицы **A**.

В неустойчивом случае (по крайней мере один из корней характеристического уравнения по модулю больше единицы) оценки (12.19) и (12.20) по-прежнему состоятельны, однако предельное распределение оценок (12.19) уже не является нормальным (см., например, [21, § 5.5]).

Если все корни характеристического уравнения по модулю больше единицы, то матрица $(\mathbf{B}_n - \mathbf{B})\,\mathbf{B}^{m-2}$ *имеет предельное распределение, которое определяется распределением случайных «возмущений»* ε_t (см. [155]). В этом утверждении $\mathbf{B}_n$ получается из **B** заменой истинных значений параметров

$\theta_1, \ldots, \theta_m$ их оценками $\widehat{\theta}_{1n}, \ldots, \widehat{\theta}_{mn}$. Приведенная формула для матрицы **A** (обратной к ковариационной матрице оценок $\widehat{\Theta}$) позволяет оценивать скорость сходимости оценок $\widehat{\Theta}$ к истинным значениям параметров Θ.

ВЫВОДЫ

1. Совокупность наблюдений $x(t_1), x(t_2), \ldots, x(t_n)$ исследуемой случайной величины, произведенных в последовательные моменты времени $t_1, t_2, \ldots, t_n$, называется *временным рядом*.

2. Временной ряд $\{x(t_1), x(t_2), \ldots, x(t_n)\}$ отличается от последовательности наблюдений $\{x_1, x_2, \ldots, x_n\}$, образующих случайную выборку, тем, что члены временного ряда *не являются ни статистически независимыми, ни одинаково распределенными.*

3. Функция $f(t)$, описывающая изменение среднего значения анализируемой переменной в зависимости от времени t ее наблюдения, называется *функцией тренда*, или просто *трендом*. Очевидно, тренд может интерпретироваться как регрессия исследуемой переменной по фактору времени[1].

4. *Моделью авторегрессии порядка m называется* модель регрессии, в которой в качестве результирующего показателя рассматривается анализируемая переменная в некоторый момент t, а в качестве объясняющих переменных (предикторов) используются значения той же самой переменной в m непосредственно предшествующих t моментов ее наблюдения. Модели авторегрессии относятся к наиболее распространенным прогностическим моделям, используемым при исследовании динамических зависимостей.

5. *Численно* задачи оценивания неизвестных значений параметров моделей авторегрессии решаются с помощью стандартного аппарата метода наименьших квадратов (см. гл. 7—9). Более сложные проблемы возникают при исследовании *статистических свойств* получаемых оценок.

---

[1]В ряде работ по математической статистике под трендом понимается лишь так называемая *долговременная составляющая* функции $\mathrm{E}x(t)$ — некое устойчивое, систематическое («очищенное» от различных периодических или сезонных колебаний) изменение в течение долгого периода (см., например, [66, с. 483]). Такое определение не претендует на математическую точность, оставаясь несколько расплывчатым: ведь понятие «долгий», «долговременный», используемое в нем, является относительным. То, что с одной точки зрения является долгим, с другой таковым не будет.

6. Оценки метода наименьших квадратов параметров модели авторегрессии в широком классе случаев (а именно при условии независимости, одинаковой распределенности и конечности дисперсий участвующих в них случайных «возмущений» ε_t, см. (12.2)) являются состоятельными. Асимптотические распределения оценок *в «устойчивом» случае* всегда являются нормальными, причем их дисперсия (ковариационная матрица) не зависит от дисперсии «возмущений» ε_t. *В общем случае* (т. е. в ситуации, когда некоторые из корней характеристического уравнения (12.17) по модулю превосходят единицу) асимптотическое распределение оценок определяется распределением случайных «возмущений» ε_t.

Математическая модель авторегрессии m-го порядка: $x_t = \sum_{\nu=1}^{m} \theta_\nu x_{t-\nu} + \varepsilon_t$, где θ_ν — параметры, подлежащие оцениванию по наблюдениям, а ε_t — случайные величины, удовлетворяющие условиям $E\varepsilon_t = 0$, $E\varepsilon_t \varepsilon_{t'} = \sigma^2$, если $t = t'$, и $t = 0$, если $t \neq t'$. В отличие от регрессионных моделей, где случайность обычно описывает погрешность наблюдений, модель авторегрессии предполагает, что уже сам исследуемый объект описывается вероятностной моделью, т. е. что анализируемая переменная x — случайная величина.

Раздел III. ИССЛЕДОВАНИЕ ЗАВИСИМОСТИ КОЛИЧЕСТВЕННОГО РЕЗУЛЬТИРУЮЩЕГО ПОКАЗАТЕЛЯ ОТ ОБЪЯСНЯЮЩИХ ПЕРЕМЕННЫХ СМЕШАННОЙ ПРИРОДЫ

Глава 13. ДИСПЕРСИОННЫЙ И КОВАРИАЦИОННЫЙ АНАЛИЗ

Допустим, что экономиста колхоза интересует зависимость урожайности какой-либо сельскохозяйственной культуры от типа почвы, сорта семян и работающего на поле звена. Допустим, далее, что в колхозе имеется I типов почвы, используется J сортов семян и обработкой культуры заняты K звеньев. Экономиста интересует: а) средняя урожайность каждого сорта семян для выбора наиболее подходящего для данного хозяйства сорта; б) влияние почвенных условий на урожайность и в) различие в трудовой отдаче звеньев, что важно для дифференцированной оплаты труда. С математической точки зрения экономист должен изучить зависимость количественной случайной величины (урожайности) от величин номинальных (сорт, звено, тип почвы). Пусть на l-м поле с i (l)-м типом почвы работало k (l)-е звено, которое возделывало j (l)-й сорт, и пусть y_l — урожайность l-го поля. Приступая к расчетам, экономист может построить математическую модель вида

$$y_l = m + a_{i(l)} + b_{j(l)} + c_{k(l)} + \varepsilon_l \quad (l = 1, \dots, L), \tag{13.1}$$

где m, a_i, b_j, c_k — константы, отражающие соответственно средний урожай в колхозе, влияние на урожай типа почвы, сорта семян и звена, а ε_l — «остаточная» случайная величина, представляющая собой отклонение наблюдаемой урожайности от модельных предположений. В модели обычно предполагается, что «остатки» ε_l независимы между собой, одинаково распределены и имеют нормальное распределение с нулевым средним. Конечно, модель (13.1) можно сделать более точной, введя в правую часть дополнительные константы, учитывающие, например, эффект взаимодействия типа почвы и сорта семян, сорта семян и звена и т. п.

Модели типа (13.1) называют моделями *дисперсионного анализа с постоянными факторами*, а совокупность методов их изучения – собственно дисперсионным анализом (ДА). Эти модели описаны в § 13.2, 13.3.

Представим теперь, что анализ проводит экономист района или области и его интересуют не успехи отдельных звеньев, а скорее тот вклад в общую изменчивость урожайности, который вносит разная работа звеньев. В этом случае постоянную $c_{k(l)}$, которая характеризовала в (13.1) работу звена, целесообразно заменить на случайную величину $\zeta(l)$ и назвать *случайным фактором*. Линейные модели ДА, содержащие только случайные факторы, называют *моделями со случайными факторами*. Модели, куда входят одновременно постоянные и случайные факторы, называют *смешанными моделями* дисперсионного анализа (последние два типа моделей описаны в § 13.4; см. также [148]).

Если требуется проанализировать данные за ряд лет, различающихся, например, по своим погодным условиям, которые характеризуются *количественными* переменными $X = (x^{(1)}, \ldots, x^{(p)})'$, то для того, чтобы учесть влияние климата, к правой части (13.1) можно добавить дополнительные члены вида $\sum_{q=0}^{m}\theta_q \cdot \psi_q(X_l)$, где θ_q — неизвестные постоянные коэффициенты, оцениваемые по обычным данным, а $\psi_q(X)$ — известные количественные (базисные) функции внешних условий X (например, средней температуры, количества осадков и т. п.). Эти дополнительные количественные факторы называют *регрессионными* переменными, а методы изучения моделей, в которых часть переменных является неколичественными, а часть количественными (регрессионными), — *ковариационным анализом* (модели ковариационного анализа посвящен § 13.5).

13.1. Классификация моделей дисперсионного анализа по способу организации исходных данных

Во введении к данному разделу описана классификация моделей, основанная на анализе математической природы входящих в них объясняющих переменных. Для дисперсионного анализа существенна также классификация, основанная на способе организации исходных данных, т. е. на том, как градации одних факторов (переменных) в исходных данных сочетаются с теми или иными градациями других переменных и как распределено общее число имеющихся наблюдений между различными возможными сочетаниями градаций переменных. Эти классификации тем более целесообразны, что ДА наиболее эффективен именно тогда, когда исследователь активно

вмешивается в организацию сбора данных или, как говорят, участвует в *планировании экспериментов*.

Предположим, что в исследование включено P факторов, причем i-й фактор имеет I_i ($i = 1, \ldots, p$) градаций, тогда имеется $\mathcal{N} = I_1 \cdot \ldots \cdot I_p$ различных сочетаний значений факторов (условий эксперимента). Если каждому из возможных условий соответствует хотя бы одно наблюдение, то такую организацию (планирование) экспериментов называют *полным P-факторным планом*. В противном случае планирование называют *неполным P-факторным планом*. В примере с колхозным экономистом для того, чтобы план был полным, необходимо, чтобы каждое из K звеньев обрабатывало бы все I типов почвы и на каждом из типов использовало бы все J сортов семян. С практической точки зрения это трудно организовать, поэтому больше распространены неполные планы.

В случаях, когда требуется сравнить в эксперименте I совокупностей условий (например, I сортов семян), часто группируют эксперименты в *блоки* (например, по типу почвы) так, чтобы внутри блока результаты эксперимента (урожай) были бы более похожи друг на друга, чем на результаты экспериментов в других блоках. Если внутри каждого из блоков удается разместить все I условий, то такой план эксперимента называют *полным блочным планом*; если только часть из них — то *неполным блочным планом*. Для того чтобы нивелировать влияние не учитываемых при анализе факторов (например, звеньев), размещение условий (сортов) внутри блока (тип почвы) часто производят случайно (по отношению к звеньям). Такие планы экспериментов называют *случайными* или *рандомизированными* планами.

13.2. Однофакторный дисперсионный анализ

13.2.1. Представление в виде регрессионной модели. Математическая модель однофакторного ДА имеет вид

$$y_{ij} = \theta_{\text{д}0} + \theta_{\text{д}i} + \varepsilon_{ij}, \quad j = 1, \ldots, J_i; \quad i = 1, \ldots, I, \qquad (13.2)$$

где $\theta_{\text{д}i}$ — неизвестные константы, удовлетворяющие равенству

$$\sum_{i=1}^{I} w_i \theta_{\text{д}i} = 0, \quad w_i \geqslant 0, \quad \sum_{i=1}^{I} w_i = 1 \qquad (13.3)$$

(w_i — некоторая заданная (вводимая исследователем) система весов)[1], а ε_{ij} — случайные погрешности, независимые между собой и имеющие нормальное распределение с нулевым средним и неизвестной дисперсией $\sigma^2 > 0$.

С содержательной точки зрения однофакторный анализ можно рассматривать как I рядов (каждый ряд длины J_i) независимых наблюдений над нормально распределенными случайными величинами со средними $\theta_{до} + \theta_{дi}$ и дисперсией σ^2.

Используя векторную запись, модель (13.2) можно представить в виде

$$\begin{bmatrix} y_{11} \\ \cdots \\ y_{1J_1} \\ y_{21} \\ \cdots \\ y_{2J_2} \\ \cdots \\ y_{I1} \\ \cdots \\ y_{IJ_I} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & 0 & 1 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & 0 & 0 & \ldots & 1 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \theta_{до} \\ \theta_{д1} \\ \cdots \\ \theta_{дI} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \cdots \\ \varepsilon_{1J_1} \\ \varepsilon_{21} \\ \cdots \\ \varepsilon_{2J_2} \\ \cdots \\ \varepsilon_{I1} \\ \cdots \\ \varepsilon_{IJ_I} \end{bmatrix} \tag{13.4}$$

или в обозначениях главы 7 (положив для краткости $n = \sum_{i=1}^{I} J_i$)

$$Y = \mathbf{X}_д \Theta_д + \boldsymbol{\varepsilon}, \tag{13.4'}$$

где Y — $(n \times 1)$-вектор наблюдений; $\mathbf{X}_д$ — $(n \times (I + 1))$-матрица плана экспериментов, имеющая ранг I; $\Theta_д$ — $((I + 1) \times 1)$-вектор неизвестных констант и $\boldsymbol{\varepsilon}$ — $(n \times 1)$-вектор случайных погрешностей[2]. Векторы Y и $\boldsymbol{\varepsilon}$ иногда обозначают так же, как $\operatorname{vec}(y_{ij})$ и $\operatorname{vec}(\varepsilon_{ij})$ или $[y_{ij}]$ и $[\varepsilon_{ij}]$. В дальнейшем мы будем пользоваться первым из этих обозна-

---

[1] Совокупность весов $\{w_i\}$ выбирают из содержательных соображений. Например, если y — урожайность, а индекс i отвечает сорту, то, если выбрать w_i пропорциональным занятой i-м сортом площади посевов в колхозе, $\theta_{до}$ будет соответствовать средней урожайности в колхозе, а $\theta_{дi}$ — разности между урожайностью i-го сорта и $\theta_{до}$.

[2] Нижний индекс $_д$ у объясняющих переменных $x^{(k)}$, матрицы плана X и оцениваемых параметров Θ подчеркивает, что речь идет об упомянутых характеристиках модели дисперсионного (а не регрессионного) анализа. Там, где это не вызывает недоразумений, этот индекс опускается.

чений. В ДА обычно проверяется гипотеза об *отсутствии* влияния рассматриваемых неколичественных переменных на результирующий показатель, т. е. $H: \theta_{д1} = \theta_{д2} = \ldots = \theta_{дI} = 0$. F-критерий для проверки этой гипотезы задается с помощью статистики (см. гл. 7).

$$F = \frac{(ОСК_H - ОСК)/(I-1)}{ОСК/(n-I)}, \qquad (13.5)$$

где $ОСК = \min\limits_{\theta_д : \Sigma w_i \theta_{дi} = 0} \sum\limits_i \sum\limits_j (y_{ij} - \theta_{до} - \theta_{дi})^2$,

а $ОСК_H = \min\limits_{\theta_{до}} \sum\limits_i \sum\limits_j (y_{ij} - \theta_{до})^2$, имеющей при правильности гипотезы H $F(I-1, n-I)$-распределение с числом степеней свободы $(I-1)$ (в числителе) и $(n-I)$ (в знаменателе). Минимизация ОСК легко выполняется методом Лагранжа. ОСК достигает минимума при $\widehat{\theta}_{до} = \Sigma w_i y_{i*}$ и $\widehat{\theta}_{дi} = y_{i*} - \widehat{\theta}_{до}$,

где $y_{i*} = \sum\limits_{j=1}^{J_i} y_{ij}/J_i$ — оценка среднего[1] в i-м ряду наблюдений; $ОСК = \sum\limits_i \sum\limits_j (y_{ij} - y_{i*})^2$; $ОСК_H$ (достигает минимума при $\widehat{\theta}_{до} = y_{**} = \Sigma y_{i*} \; J_i/n$ и равно $\sum\limits_i \sum\limits_j (y_{ij} - y_{**})^2$. Путем несложных алгебраических преобразований получаем

$$\sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - y_{**})^2 = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - y_{i*} + y_{i*} - y_{**})^2 =$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{ij} - y_{i*})^2 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} (y_{i*} - y_{**})^2,$$

так как сумма со смешанными произведениями равна нулю. Откуда $ОСК_H - ОСК = \sum\limits_{i=1}^{I} J_i (y_{i*} - y_{**})^2$, и критерий F принимает вид

$$F = \frac{\sum\limits_{i=1}^{I} J_i (y_{i*} - y_{**})^2/(I-1)}{\sum\limits_{i=1}^{I} \sum\limits_{j=1}^{J_i} (y_{ij} - y_{i*})^2/(n-I)} \qquad (13.5')$$

---

[1]Подстрочный индекс $_*$ означает усреднение по индексу, который он заменяет.

Числитель F обозначают s_H^2, а знаменатель — s_e^2. Таким образом, если окажется, что подсчитанная по формуле (13.5') (или (13.5)) величина F превосходит значение 100α%-ной точки F_α $(I-1, n-I)$ F-распределения с числом степеней свободы числителя, равным $I-1$, и знаменателя — $n-I$ (см. табл. П.5), то гипотеза H отвергается (с уровнем значимости критерия, равным α). Различные суммы квадратов, встречающиеся в ДА, принято располагать в виде специальной таблицы ДА для однофакторного анализа (табл. 13.1). Последний столбец таблицы объяснен в следующем пункте.

Таблица 13.1

Источник изменчивости	Сумма квадратов (СК)	Число степеней свободы (чсс)	$\overline{СК}$ = СК/чсс	E ($\overline{СК}$)
Между градациями	$СК_H=\sum_i J_i(y_{i*}-y_{**})^2$	$I-1$	s_H^2	$\sigma^2+(I-1)^{-1}\times\sum_i J_i(\theta_i-\theta_*)^2$
Ошибки	$СК_e=\sum_i\sum_j(y_{ij}-y_{i*})^2$	$\sum_i J_i-I$	s_e^2	σ^2
«Полная» сумма квадратов	$СК_П=\sum_i\sum_j(y_{ij}-y_{**})^2$	$\sum_i J_i-1$	—	

Терминология для сумм, используемых в столбце «источник изменчивости», в разных работах разная. Так, вместо термина «между градациями» употребляют термины «между совокупностями», «между способами обработки»; вместо термина «ошибка» говорят о сумме квадратов «внутри групп», «внутри совокупностей», «остаточной» сумме квадратов.

13.2.2. Геометрический смысл ДА. Хотя общие вопросы проверки гипотез в случае линейной регрессии уже рассмотрены в гл. 7, представляется интересным конкретизировать их в случае однофакторной модели ДА. Положим

$$\varepsilon_{i*}=\sum_j \varepsilon_{ij}/J_i;\quad \varepsilon_{**}=\sum_i J_i\,\varepsilon_{i*}/n;\quad \theta_*=\sum_i J_i\,\theta_i/n$$

и определим n-мерные векторы аналогично тому, как это было сделано в предыдущем пункте:

$$U=\mathrm{vec}\,(y_{**}-\theta_*);\quad V=\mathrm{vec}\,(y_{i*}-y_{**});\quad W=\mathrm{vec}\,(y_{ij}-y_{i*}).$$

Из модели (13.2) следует, что $U = \mathrm{vec}(\varepsilon_{**})$, $W = \mathrm{vec}(\varepsilon_{ij} - \varepsilon_{i*})$, а

$$V = V_0 + \mathrm{vec}(\theta_i - \theta_*), \tag{13.6}$$

где $V_0 = \mathrm{vec}(\varepsilon_{i*} - \varepsilon_{**})$.

Из тождества

$$\varepsilon_{ij} = \varepsilon_{**} + (\varepsilon_{i*} - \varepsilon_{**}) + (\varepsilon_{ij} - \varepsilon_{i*})$$

следует, что $\varepsilon = U + V_0 + W$.

Векторы U, V_0, W взаимно ортогональны, что легко проверяется непосредственно, поэтому

$$\varepsilon'\varepsilon = U'U + V_0'V_0 + W'W. \tag{13.7}$$

Квадратичные формы, стоящие в правой части (13.7), взаимно независимы и имеют ранги, в силу определения и условия (13.3), соответственно равные 1, $I - 1$, $n - I$. Поскольку ранг правой части (13.7) равен $n = 1 + (I - 1) + (n - I)$, в силу теоремы Кохрана ([148, приложение VI]), отсюда следует, что $U'U/\sigma^2$, $V_0'V_0/\sigma^2$, $W'W/\sigma^2$ имеют χ^2-распределения с числами степеней свободы, равными их рангам. Таким образом, числитель и знаменатель критерия (13.5), (13.5′) независимы, и F имеет F-распределение с числами степеней свободы $I - 1$, $n - I$. В случае, когда H не имеет места, F имеет нецентральное F-распределение с тем же числом степеней свободы и параметром нецентральности δ^2, равным в силу (13.6)

$$\delta^2 = \mathrm{vec}(\theta_i - \theta_*)' \,\mathrm{vec}(\theta_i - \theta_*)/\sigma^2 = \sum_i J_i (\theta_i - \theta_*)^2/\sigma^2. \tag{13.8}$$

Диаграммы для нахождения мощности F-критерия при заданных n, I, δ можно найти в [148].

13.2.3. Доверительные интервалы. Если в результате применения F-критерия гипотеза H отвергается, то следующий шаг состоит в выяснении того, насколько параметры θ_i отличаются друг от друга. В частности, обычно представляют интерес разности вида $\theta_1 - \theta_2$, $\theta_1 - (\theta_2 + \theta_3)/2$, $(\theta_1 + \theta_2)/2 - (\theta_3 + \theta_4 + \theta_5)/3$ и т. п. Эти линейные комбинации, имеющие вид $\Sigma c_i\theta_i = 0$, где $\Sigma c_i = 0$, называются *сравнениями* или *контрастами* (contrast) параметров θ_i. Если бы линейная комбинация была *задана до получения экспериментальных данных*, то $(1 - \alpha)$ — доверительный интервал для $\Sigma c_i\theta_i$ — мы могли бы построить как

$$\Sigma c_i y_{i*} \pm (t_{\alpha/2}(n - I)) \,\Sigma \frac{c_i^2}{J_i} s_e, \tag{13.9}$$

где $t_\beta(k)$ — $100\beta\%$-ная точка t-распределения Стьюдента с k степенями свободы.

Однако на практике представляющие интерес сравнения составляются обычно *после получения экспериментальных данных*, т. е. тогда, когда уже известны оценки $\widehat{\theta}_i$. Исследователь, опираясь на них, среди всех возможных сравнений отбирает те, которые кажутся ему наиболее важными. Применение формулы (13.9) к отобранным сравнениям не оправдано и приводит к более узкому, чем должно быть, доверительному интервалу. Тактика исследователя в этих условиях должна заключаться в том, чтобы отказавшись от индивидуального доверительного интервала строить доверительные интервалы *множественные*, которые *одновременно* выполнялись бы либо *для всех возможных сравнений*, либо *для* какого-либо *выделенного подмножества сравнений*. Наиболее известны три метода построения таких интервалов: S-метод Шеффе, Т-метод Тьюки и метод уменьшения уровня критерия Стьюдента.

S-метод Шеффе опирается на следующее простое рассуждение:

$$\left|\sum_i c_i(\theta_i - y_{i*})\right| = \left|\sum c_i(\varepsilon_{i*} - \varepsilon_{**})\right| =$$

$$= \left|\sum_i \frac{c_i}{\sqrt{J_i}}\sqrt{J_i}(\varepsilon_{i*} - \varepsilon_{**})\right| \leqslant \left(\sum_i \frac{c_i^2}{J_i}\right)^{1/2} \times$$

$$\times \left(\sum_k J_k(\varepsilon_{k*} - \varepsilon_{**})^2\right)^{1/2}. \qquad (13.10)$$

Правая часть (13.10) состоит из двух сомножителей, первый из которых носит неслучайный характер, а второй не зависит от выбора c_i, распределен как $\sqrt{\sigma^2\chi^2(I-1)}$ и не зависит от s_e^2. Отсюда можно вывести, что величина второго сомножителя с вероятностью $(1-\alpha)$ будет меньше, чем $[(I-1) \times F_\alpha(I-1, n-1)]^{1/2}s_e$. Следовательно, с вероятностью не меньшей $1-\alpha$, для всех сравнений одновременно выполняется неравенство

$$|\Sigma c_i\theta_i - \Sigma c_i y_{i*}| \leqslant \left(\sum_i \frac{c_i^2}{J_i}\right)^{1/2} \times$$

$$\times [(I-1)F_\alpha(I-1,\ n-I)]^{1/2}\, s_e. \qquad (13.10')$$

Т-метод Тьюки применяется только к сравнениям вида $\theta_i - \theta_j$. Пусть $y_{i*} - \theta_i$ расположены в вариационный ряд, обозначим $Z_{\min}$ — наименьшее из них и $Z_{\max}$ — наибольшее. Для всех $I(I-1)/2$ пар (i, j)

$$|y_{i*} - y_{j*} - \theta_i + \theta_j| \leqslant Z_{\max} - Z_{\min} =$$

$$= \max_i \varepsilon_{i*} - \min_j \varepsilon_{j*}. \qquad (13.11)$$

Разность в правой части неравенства (13.11) при $J_i = J$ с вероятностью $1 - \alpha$ ограничена величиной $q_\alpha(I, n - I)s_e$, где $q_\alpha(I, n - I)$ — 100α%-ная точка стьюдентизированного размаха с числом степеней свободы I, $n - I$[1].

Метод уменьшения уровня критерия Стьюдента. Если требуется построить k доверительных интервалов, где k не слишком велико, то можно воспользоваться неравенством (13.9) с меньшим значением уровня $\alpha' = \alpha/k$. В этом случае вероятность того, что будут верны одновременно все k доверительных интервалов, не менее $1 - \alpha$.

Пример 13.1. Допустим, что $\alpha = 0{,}05$, $I = 5$, $J_i = 6$, $i = 1, \ldots, 5$. Доверительные интервалы строятся только для разностей $\theta_i - \theta_j$ ($k = 10$). Тогда введенные выше три метода дают следующую длину доверительных интервалов (в единицах s_e).

$$S\text{-метод: } 2\left(\Sigma \frac{c_i^2}{J_i}\right)^{1/2} [(I-1)(F_\alpha(I, n-I)]^{1/2} = 11{,}50;$$

$$T\text{-метод: } 2q_\alpha(I, n-I) = 8{,}32;$$

метод уменьшения уровня критерия Стьюдента в k раз:

$$2\left(\Sigma \frac{c_i^2}{J_i}\right)^{1/2} t_{\alpha/(2k)}(n-I) = 10{,}67.$$

Таким образом, наименее экономным оказался S-метод, но это и естественно, так как в нем интервал рассчитан на произвольное сравнение.

---

[1]Пусть случайные величины $\eta_1, \ldots, \eta_I$, $\chi^2(k)$ независимы между собой и $\eta_i \in N(0, 1)$, тогда случайная величина $q_{Ik} = (\max \eta_i - \min \eta_i)/\sqrt{\chi^2(k)}$ называется *стьюдентизированным размахом* с числом степеней свободы I, k.

13.3. Полный двухфакторный дисперсионный анализ

13.3.1. Взаимодействия. Рассмотрим полный двухфакторный эксперимент с факторами A и B, имеющими соответственно I и J градаций (уровней). Обозначим θ_{ij} среднее значение результата эксперимента при сочетании i-го уровня фактора A с j-м уровнем фактора B (среднее значение в (i, j)-ячейке прямоугольной таблицы, в которой строкам соответствуют градации фактора A, а столбцам — градации фактора B). Пусть $\{v_i\}$, $\{w_i\}$ $(\sum\limits_i v_i = \sum\limits_j w_j = 1,\ v_i \geqslant 0,\ w_j \geqslant 0)$ системы весов, выбранные в соответствии с градациями факторов A и B. Например, если фактор A — тип почвы, а B — сорт, результат эксперимента — урожай, то v_i может быть пропорционально площадям с i-м типом почвы, а w_j — площадям, занятым j-м сортом. *Средним* значением i-го *уровня* (*фактора*) A называют $a_i = \sum\limits_j w_j\theta_{ij}$.

В рассмотренном выше примере a_i — ожидаемая средняя урожайность на i-м типе почвы при условии, что выбор сорта при посеве не зависит от типа почвы и пропорционален $\{w_j\}$. Эти условия выполняются, в частности, когда либо тип почвы оказывает приблизительно одинаковое влияние на урожайность всех сортов, либо низка общая культура земледелия, и землепользователю просто неизвестны или он лишен возможности использовать оптимальные сочетания сорта и типа почвы. Аналогично определяется среднее j-го уровня (фактора) B $b_j = \sum\limits_i v_i\theta_{ij}$. В примере b_j — это ожидаемая средняя урожайность j-го сорта при случайном распределении его по полям, пропорциональном $\{v_i\}$.

Генеральным средним называют

$$\theta_0 = \sum_j w_j b_j = \sum_i v_i a_i = \sum_i \sum_j v_i w_j \theta_{ij}.$$

Главный эффект i-го уровня A определяется как $\alpha_i = a_i - \theta_0$. Очевидно, $\{\alpha_i\}$ удовлетворяют условию

$$\sum_i v_i \alpha_i = 0. \tag{13.12}$$

Аналогично главный эффект j-го уровня B определяется как $\beta_j = b_j - \theta_0$, и $\{\beta_j\}$ удовлетворяют условию

$$\sum_j w_j \beta_j = 0. \tag{13.13}$$

Взаимодействием i-го уровня A с j-м уровнем B называют

$$\gamma_{ij}=\theta_{ij}-a_i-b_j+\theta_0=(\theta_{ij}-\theta_0)-\alpha_i-\beta_j.$$

Взаимодействия удовлетворяют условиям:

$$\sum_i v_i\gamma_{ij}=0 \quad \text{для всех } j=1,\ldots,J;$$
$$\sum_j w_j\gamma_{ij}=0 \quad \text{для всех } i=1,\ldots,I. \qquad (13.14)$$

Итак, математическая модель для средних имеет вид

$$\theta_{ij}=\theta_0+\alpha_i+\beta_j+\gamma_{ij}, \qquad (13.15)$$

где α_i, β_j, γ_{ij} удовлетворяют уравнениям (13.12)—(13.14).

Очень важным для практического использования ДА является следующее утверждение [148, § 4.1]:
Если при некоторой системе весов $\{v_i\}$, $\{w_j\}$ *все взаимодействия равны нулю, то они равны нулю при любой другой системе весов.* В этом случае модель называют *аддитивной. В аддитивной модели каждое сравнение средних* $\{a_i\}$ *или* $\{b_j\}$ *и главных эффектов* $\{\alpha_i\}$ *или* $\{\beta_j\}$ *имеет значение, не зависящее от системы весов* $\{v_i\}$ *и* $\{w_j\}$.

Содержательная интерпретация результатов ДА проста и наглядна, когда мы решаем (на основании статистических или других соображений), что взаимодействия отсутствуют. В этом случае заключением о главных эффектах можно завершить весь анализ. Однако если гипотеза об отсутствии взаимодействий была принята на основании только того, что она не была отвергнута некоторым F-критерием, то следует посмотреть мощность этого критерия и оценить, насколько статистически невыявленные взаимодействия могут повлиять на интерпретацию результатов.

Часто величины взаимодействий могут быть уменьшены с помощью подходящего преобразования наблюдаемых случайных величин [14, п. 10.3.4]. Пример использования преобразования с целью сделать модель аддитивной можно найти в [170].

Среди статистиков нет единого мнения о целесообразности изучения главных эффектов в условиях статистически значимых взаимодействий. Нам кажется, что этот вопрос носит, скорее, содержательный, чем формально-математический характер. Если введенные средние имеют смысл, то изучение их свойств целесообразно. Наконец, взаимодействия могут быть

значимы статистически, но относительно малы по сравнению с главными эффектами.

Всюду в дальнейшем, если не оговорено противное, будет выбираться система весов: $v_i = 1/I$, $w_j = 1/J$. Такие веса иногда называют *равными*.

13.3.2. Двухфакторный анализ с равным числом K наблюдений в ячейках ($K > 1$). Обозначим y_{ijk} k-е наблюдение в (i, j)-ячейке, тогда математическая модель имеет вид

$$y_{ijk} = \theta_0 + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ i = \overline{1, I};\ j = \overline{1, J},\ k = \overline{1, K}, \quad (13.16)$$

где ε_{ijk} независимы и имеют нормальное распределение со средним, равным нулю, и дисперсией σ^2, а константы α_i, β_j, γ_{ij} удовлетворяют соотношениям (13.12)—(13.14) с равными весами. Обычно проверяются следующие гипотезы:

H_A: все $\alpha_i = 0$;

H_B: все $\beta_j = 0$;

H_{AB}: все $\gamma_{ij} = 0$.

Оценки для параметров строятся аналогично тому, как это делается в § 13.2.1:

$$\widehat{\theta}_0 = y_{***},\ \widehat{\alpha}_i = y_{i**} - y_{***},\ \widehat{\beta}_j = y_{*j*} - y_{***},$$

$$\widehat{\gamma}_{ij} = y_{ij*} - y_{i**} - y_{*j*} + y_{***}.$$

Возьмем тождество

$$y_{ijk} - y_{***} = (y_{ijk} - y_{ij*}) + (y_{ij*} - y_{i**} - y_{*j*} + y_{***}) + \\ + (y_{i**} - y_{***}) + (y_{*j*} - y_{***}),$$

возведем в квадрат его правую и левую части и просуммируем по всем значениям индексов i, j, k. Так как все смешанные произведения при суммировании равны нулю, получаем

$$\text{СК}_\text{П} = \text{СК}_e + \text{СК}_{AB} + \text{СК}_A + \text{СК}_B. \quad (13.17)$$

Поскольку сумма рангов слагаемых правой части (13.17) совпадает с рангом квадратичной формы левой части, согласно теореме Кохрана получаем, что суммы квадратов правой части независимы и соответственно распределены как $\sigma^2 \cdot \chi^2$ с числом степеней свободы, совпадающим с рангом соответствующей формы. Таким образом, критерии для проверки гипотез H_A, H_B, H_{AB} могут быть построены как соответствующие F-отношения s_A^2/s_e^2, s_B^2/s_e^2, s_{AB}^2/s_e^2, где s_A^2, s_B^2, s_{AB}^2, s_e^2 определены в таблице ДА для двухфакторного анализа с $K > 1$ наблюдениями в ячейке (табл. 13.2).

Таблица 13.2

Источник изменчивости	Сумма квадратов (СК)	Число степеней свободы (чсс)	$\overline{СК} = \frac{СК}{чсс}$	E ($\overline{СК}$)
Главные эффекты A	$СК_A = JK \sum_i (y_{i**} - y_{***})^2$	$I-1$	s_A^2	$\sigma^2 + JK(I-1)^{-1} \sum_i \alpha_i^2$
Главные эффекты B	$СК_B = IK \sum_j (y_{*j*} - y_{***})^2$	$J-1$	s_B^2	$\sigma^2 + IK(J-1)^{-1} \sum_j \beta_j^2$
Взаимодействия AB	$СК_{AB} = K \sum_i \sum_j (y_{ij*} - y_{i**} - y_{*j*} + y_{***})^2$	$(I-1)(J-1)$	s_{AB}^2	$\sigma^2 + K(I-1)^{-1}(J-1)^{-1} \sum_i \sum_j \gamma_{ij}^2$
Ошибки	$СК_e = \sum_i \sum_j \sum_k (y_{ijk} - y_{ij*})^2$	$IJ(K-1)$	s_e^2	σ^2
«Полная» сумма квадратов	$СК_П = \sum_i \sum_j \sum_k (y_{ijk} - y_{***})^2$	$IJK-1$	—	—

Доверительные интервалы для сравнений $\{\alpha_i\}$ $\{\beta_j\}$, $\{\gamma_{ij}\}$ могут быть построены с помощью S-метода аналогично тому, как это сделано в случае однофакторного анализа. T-метод может быть также применен к разностям главных эффектов. Однако к взаимодействиям он уже не применим.

13.3.3. Случай неравных K_{ij}. В общем случае не существует простого разложения, подобного (13.17). Поэтому остаются две возможности: либо воспользоваться точным, но громоздким методом анализа, изложенным, например, в монографии [148], либо применить приближенный метод. Последний заключается в следующем. Приближенно считают, что

$$y_{ij*} \approx \theta_0 + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij},$$

где ε_{ij} независимы и нормально распределены с нулевым средним и дисперсией $\sigma^2/\overline{K}$; $\overline{K} = \sum_i\sum_j K_{ij}/IJ$, а $\{\alpha_i\}$, $\{\beta_j\}$, $\{\gamma_{ij}\}$ удовлетворяют соотношениям (13.12)—(13.14) с $v_i = 1/I$, $w_j = 1/J$. Оценка σ^2 в этом случае строится с помощью $СК_e = \sum_i\sum_j\sum_k (y_{ijk} - y_{ij*})^2$. Далее анализ проводится так же, как в п. 13.3.2 с $K = \overline{K}$.

В частном случае, когда для всех (i, j) $K_{ij} = K_{i\cdot}K_{\cdot j}/K_{\cdot\cdot}$, где $K_{i\cdot} = \sum_i K_{ij}$, $K_{\cdot j} = \sum_i K_{ij}$, $K_{\cdot\cdot} = \sum_i K_{i\cdot} = \sum_i K_{\cdot j}$, при выборе $v_i = K_{i\cdot}/K_{\cdot\cdot}$ и $w_j = K_{\cdot j}/K_{\cdot\cdot}$ разложение типа (13.17) имеет место, и ДА проводится с очевидными изменениями в табл. 13.2.

13.3.4. Случай $K_{ij} = 1$. Как правило, подобные данные возникают при использовании планов с рандомизированными блоками, когда можно априори ожидать, что или взаимодействие между изучаемым фактором A (способом обработки) и фактором B (блоком) мало в среднем, или что существует малопараметрическая параметризация взаимодействий. Ниже указываются основные модели.

Аддитивная модель без взаимодействий

$$y_{ij} = \theta + \alpha_i + \beta_j + \varepsilon_{ij},\quad i = 1, \ldots, I;\ j = 1, \ldots, J, \tag{13.18}$$

где $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, а ε_{ij} независимы, нормально распределены $N(0, \sigma^2)$. При рандомизированных блоках ε_{ij} отражает суммарное влияние двух источников случайных отклонений: отклонения, связанного со случайностью расположения фактора A в блоке, и отклонения, равного случайной погрешности воспроизведения эксперимента при заданном расположении A в блоке. Поскольку в аддитивной модели предполагается, что взаимодействия отсутствуют, проверка гипотез $H_A : \alpha_i = 0$ $(i = 1, \ldots, I)$ и $H_B : \beta_j = 0$ $(j = 1, \ldots, J)$ проводится так же,

как в п. 13.3.2, только роль суммы квадратов ошибок играет сумма квадратов взаимодействий. Табл. 13.3 — это таблица ДА для аддитивной модели двухфакторного анализа с одним наблюдением в ячейке.

Таблица 13.3

Источник изменчивости	Сумма квадратов (СК)	Число степеней свободы (чсс)	$\overline{СК}=\frac{СК}{чсс}$	$E(\overline{СК})$
Главные эффекты (способы обработки)	$СК_A=J\sum_i (y_{i*}-y_{**})^2$	$I-1$	s_A^2	$\sigma^2+J\times$ $\times(I-1)^{-1}\Sigma\alpha_i^2$
Главные эффекты (блоки)	$СК_B=I\sum_j (y_{*j}-y_{**})^2$	$J-1$	s_B^2	$\sigma^2+I\times$ $\times(J-1)^{-1}\Sigma\beta_j^2$
Ошибки	$СК_e=\sum_i\sum_j (y_{ij}-$ $-y_{i*}-y_{*j}+y_{**})^2$	$(I-1)\times$ $\times(J-1)$	s_e^2	σ^2
«Полная» сумма квадратов	$СК_п=\sum_i\sum_j (y_{ij}-y_{**})^2$	—	—	—

Модель с однопараметрическим описанием взаимодействий носит стандартный вид (13.16), только дополнительно предполагается, что

$$\gamma_{ij}=\gamma\alpha_i\beta_j. \tag{13.19}$$

Гипотеза об отсутствии взаимодействий в этом случае эквивалентна гипотезе $H_Г:\gamma=0$. Для ее проверки можно использовать критерий

$$F=\frac{СК_Г}{(ОСК-СК_Г)/[(I-1)(J-1)-1]}, \tag{13.20}$$

где

$$СК_Г=\frac{\left(\sum_i\sum_j\widehat{\alpha}_i\widehat{\beta}_i y_{ij}\right)^2}{\sum_i\widehat{\alpha}_i^2\sum_j\widehat{\beta}_j^2};\ ОСК=\sum_i\sum_j(\widehat{\gamma}_{ij})^2; \tag{13.21}$$

$$\widehat{\alpha}_i=y_{i*}-y_{**};\ \widehat{\beta}_j=y_{*j}-y_{**};\ \widehat{\gamma}_{ij}=y_{ij}-y_{i*}-y_{*j}+y_{**}.$$

Критерий F в (13.20) при $\gamma = 0$ имеет F-распределение с 1, $(IJ - I - J)$ степенями свободы. Гипотезы H_A и H_B проверяются так же, как в п. 13.3.2, только сумма квадратов ошибок определяется как $СК_e = ОСК - СК_Г$ и имеет на одну степень свободы меньше, чем в табл. 13.3.

Более реалистической является *модель с $I + J - 1$ параметрическим описанием взаимодействий* [217]. Она похожа на предыдущую, только в ней предполагается, что взаимодействия пропорциональны произведению не главных факторов, а некоторых констант, оцениваемых по выборке,

$$\gamma_{ij} = \lambda f_i g_j, \text{ где } \sum_i f_i = \sum_j g_j = 0, \ \sum_i f_i^2 = \sum_j g_j^2 = 1.$$

Пусть $\Gamma = \|\widehat{\gamma}_{ij}\|$ — $(I \times J)$-матрица с $\widehat{\gamma}_{ij}$, определенными, как в (13.21), тогда $\widehat{\alpha}_i$ и $\widehat{\beta}_j$ определяются, как выше, а $\widehat{\lambda}^2 = l_1$, где l_1 — наибольшее собственное число матрицы $\Gamma'\Gamma$; $\widehat{F} = \{\widehat{f}_i\}'$ — нормализованный собственный вектор $\Gamma\Gamma'$, соответствующий собственному числу l_1; $\widehat{G} = \{\widehat{g}_j\}'$ — аналогичный вектор $\Gamma'\Gamma$. Для проверки гипотезы $H_0 : \lambda = 0$ используется отношение типа (13.20) с естественной заменой $СК_Г$ на l_1. Однако распределение отношения сложнее и не сводится к F-распределению.

13.4. Модели дисперсионного анализа со случайными факторами

13.4.1. Место моделей со случайными факторами. В ряде экспериментов, проводимых по схеме дисперсионного анализа, значения, которые принимает некоторый фактор, нельзя охарактеризовать ничем, кроме того, что они «наудачу» извлечены из некоторой генеральной совокупности. В этом случае вряд ли целесообразно связывать с индивидуальными значениями фактора числовые характеристики их вклада в общий итог (y), а лучше оценить вклад фактора в целом в общую изменчивость y. В этом случае, как указано в § 13.1, и возникают модели со случайными факторами. При этом, если окажется, что вклад фактора в общую изменчивость достаточно велик, то целесообразно из профессиональных содержательных соображений найти характеристики значений фактора, предположительно связанные с y, и изучить с помощью традиционного анализа их влияние на итог. Так, если нас интересует влияние личности рабочего на его производительность труда, то здесь

целесообразна случайная выборка среди всех рабочих, занятых однотипным трудом, и использование модели со случайным фактором — влиянием личности рабочего. Если же нас интересует различие в производительности труда рабочих с различным уровнем образования, то целесообразно разбить рабочих на группы по уровню образования и взять для сравнения по нескольку человек из каждой группы. Здесь больше подходит смешанная модель вида

$$y_{ijk} = \theta_0 + \theta_i + \zeta_{ij} + \varepsilon_{ijk}, \qquad (13.22)$$

где y_{ijk} — производительность в k-й день j-го рабочего с i-м уровнем образования; θ_0 — среднее значение производительности труда; θ_i — постоянная, отражающая влияние i-го уровня образования ($\Sigma\theta_i = 0$); ζ_{ij} — случайный фактор, характеризующий отклонение от общего среднего производительности труда j-го рабочего из группы с уровнем образования i; ε_{ijk} — случайное колебание в производительности труда в k-й день эксперимента этого рабочего.

Другой пример. При изучении воспроизводимости химических анализов часто говорят о средней межлабораторной ошибке [95]. В этом термине различные лаборатории рассматриваются как случайные, наудачу выбранные из совокупности лабораторий какой-либо отрасли. Для оценки этой ошибки подходит модель со случайным фактором — названием лаборатории. Если же лаборатории можно классифицировать по какому-нибудь доступному и существенному для точности анализов признаку, например по уровню оснащения современным оборудованием, то межлабораторную ошибку целесообразно определять отдельно внутри однородных по этому признаку классов. В частности, можно использовать модель вида (13.22), в которой первый фактор (i) соответствует уровню оснащения лаборатории современным оборудованием, а второй фактор (j) отражает условный номер лаборатории среди однотипных по оборудованию лабораторий.

В модели (13.22) мы встретились с так называемым *иерархическим* или *группированным*, или, как еще говорят, *гнездовым планом*, в котором каждому уровню первого фактора (i) соответствует свое подмножество значений второго фактора (j).

13.4.2. Однофакторный анализ. Простейшая математическая модель имеет вид

$$y_{ij} = \theta + \zeta_i + \varepsilon_{ij}, \quad i = 1,\ldots, I; \; j = 1,\ldots, J, \qquad (13.23)$$

где θ — постоянная; ζ_i — случайная составляющая, отражающая влияние i-го выборочного значения фактора ζ; ε_{ij} —

случайная ошибка; $I + J \cdot I$ случайные величины $\{\zeta_i\}$ и $\{\varepsilon_{ij}\}$ независимы в совокупности и $\zeta \in N(0, \sigma_\zeta^2)$, $\varepsilon \in N(0, \sigma_e^2)$. Поскольку $\sigma_y^2 = \sigma_\zeta^2 + \sigma_e^2$, величины σ_ζ^2 и σ_e^2 называют компонентами дисперсии (наблюдения).

Модель (13.23) существенно отличается от близкой модели с постоянными факторами (13.2). В ней: 1) все наблюдения имеют одинаковые математические ожидания и 2) наблюдения не являются статистически независимыми. В частности, положительно коррелированы наблюдения y_{ij} и $y_{ij'}$ $(j \neq j')$, соответствующие одному и тому же значению (i) изучаемого фактора. Соответствующий коэффициент корреляции, равный $\rho = \sigma_\zeta^2/(\sigma_\zeta^2 + \sigma_e^2)$, называют *коэффициентом внутриклассовой корреляции.*

Основная проверяемая гипотеза

$$H: \sigma_\zeta^2 \leqslant \lambda_0 \sigma_e^2, \tag{13.24}$$

где $\lambda_0 \geqslant 0$ — заданная константа. Определим s_H^2 и s_e^2 так, как указано в табл. 13.1, и построим F-отношение (13.5′). Статистический критерий для проверки гипотезы (13.24) $F \leqslant F_\alpha(I-1, (J-1)I)$ имеет мощность, выражающуюся только через центральное F-распределение [148],

$$\beta(\lambda) = P\left\{F(I-1, I(J-1)) \geqslant F_\alpha(I-1, I(J-1)) \times \times \frac{1+J\lambda_0}{1+J\lambda}\right\}. \tag{13.25}$$

Точечные оценки для компонент дисперсии

$$\widehat{\sigma_e^2} = s_e^2 \text{ и } \widehat{\sigma_\zeta^2} = J^{-1}(s_H^2 - s_e^2). \tag{13.26}$$

Одна из трудностей дисперсионного анализа со случайными факторами состоит в том, что для негауссовских распределений среднее арифметическое и среднеквадратическое отклонения уже не являются устойчивыми оценками параметров положения и масштаба [14, п. 10.4.4]. Для того чтобы обойти эту трудность, в однофакторном случае может быть рекомендована следующая приближенная процедура:

1) для каждой (по i) серии наблюдений y_{ij} $(j = 1, \ldots, J)$ вычислить устойчивые оценки положения и масштаба. Обозначить их $\tilde{y}_{i*}$ и $\tilde{s}_i$;

2) заменить исходные серии наблюдений y_{ij} на

$$z_{ik} = \tilde{y}_{i*} + \tilde{s}_i u_{k/(J+1)} \quad (k = 1, 2, \ldots, J);$$

(здесь u_q — q-квантиль стандартного $N(0, 1)$-распределения).

3) провести с наблюдениями z_{ik} обычный дисперсионный анализ по описанным выше формулам.

13.4.3. Иерархический план на двух уровнях. В качестве примера рассмотрим сначала описанную в п. 13.4.1 задачу по определению точности проведения химического анализа в лабораториях какой-либо отрасли промышленности. Предположим, что все лаборатории отрасли могут быть разбиты на группы, примыкающие к городам, в которых есть метрологические центры по данному виду анализа. Эксперимент состоит в том, что сначала наудачу выбирается несколько метрологических центров (городов), а затем для каждого из выбранных центров также наудачу отбирается несколько из примыкающих к нему лабораторий, и в каждой из лабораторий проводится несколько повторных определений одного и того же образца. Простейшая математическая модель для рассматриваемого иерархического плана с двумя случайными факторами имеет вид

$$y_{ijk} = \theta + \zeta_i + \eta_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, I; \; j = 1, \ldots, J;$$
$$k = 1, \ldots, K, \tag{13.27}$$

где y_{ijk} — результат k-го анализа в j-й лаборатории, примыкающей к i-му городу; θ — истинная (обычно неизвестная) определяемая концентрация; ζ_i — эффект i-го метрологического центра; η_{ij} — эффект j-й лаборатории i-го центра и ε_{ijk} — случайная ошибка k-го повторения анализа в j-й лаборатории i-го центра; величины $\{\zeta_i\}$, $\{\eta_{ij}\}$, $\{\varepsilon_{ijk}\}$ взаимно независимы и нормально распределены с нулевыми средними и дисперсиями σ_ζ^2, σ_η^2 и σ_e^2 соответственно.

Разложение общей суммы квадратов на части, соответствующие рассматриваемой модели, показано в таблице ДА для иерархического плана с двумя случайными факторами (табл. 13.4).

F-критерии для гипотез $H_\zeta : \sigma_\zeta^2 = 0$, $H_\eta : \sigma_\eta^2 = 0$ строятся с помощью соответствующих отношений средних квадратов. Например, для проверки H_ζ используется отношение s_ζ^2/s_η^2, которое в случае, когда гипотеза H_ζ верна, имеет $F(I-1, I(J-1))$ распределение. В рассматриваемом случае имеет смысл и оценка суммы $\sigma_y^2 = \sigma_\zeta^2 + \sigma_\eta^2 + \sigma_e^2$, которая является характеристикой воспроизводимости анализа в случайно выбранной лаборатории.

В заключение рассмотрим иерархическую смешанную модель (13.22). В дополнение к сделанным выше (п. 13.4.1) предположениям допустим, что $\{\zeta_i\}$ и $\{\varepsilon_{ij}\}$ взаимно независимы,

Таблица 13.4

№ п/п	Источник изменчивости	Сумма квадратов (СК)	Число степеней свободы (чсс)	$\overline{СК} = \frac{СК}{чсс}$	$E(\overline{СК})$
1	Между градациями первого фактора (ζ)	$KJ\sum_i (y_{i**}-y_{***})^2$	$I-1$	s_ζ^2	$\sigma_e^2+K\sigma_\eta^2+$ $+KJ\sigma_\zeta^2$
2	Между градациями второго фактора (η) по первому фактору (ζ)	$K\sum_i\sum_j (y_{ij*}-y_{i**})^2$	$I(J-1)$	s_η^2	$\sigma_e^2+K\sigma_\eta^2$
3	Ошибка	$\sum_i\sum_j\sum_k (y_{ijk}-y_{***})^2$	$IJ(K-1)$	s_e^2	σ_e^2

$\zeta_i \in N(0, \sigma_\zeta^2)$ и $\varepsilon_{ij} \in N(0, \sigma_e^2)$. В этом случае таблица дисперсионного анализа для модели (13.22) получается из табл. 13.4 путем замены в первой строке s_ζ^2 на s_θ^2 и $\sigma_e^2+K\sigma_\eta^2+KJ\sigma_\zeta^2$ на $\sigma_e^2+K\sigma_\eta^2+(I-1)^{-1}KJ\sum_i\theta_i^2$

13.5. Ковариационный анализ (КА) и проблема статистического исследования смесей многомерных распределений

13.5.1. Определение и модель ковариационного анализа. Следуя [6], определим ковариационный анализ (КА) как совокупность методов и результатов, относящихся к математико-статистическому анализу моделей, предназначенных для исследования зависимости среднего значения некоторого количественного результирующего показателя y от набора неколичественных факторов $X_д$ и одновременно от набора количественных (*регрессионных* или *сопутствующих*) переменных X. Результирующий признак y может быть векторным (тогда говорят о *многомерном* ковариационном анализе).

Неколичественные факторы $X_д$ задают сочетания условий (качественной природы), в которых производилась фиксация каждого из наблюдений (экспериментальных значений) y и X, и описываются обычно с помощью так называемых *индикаторных* переменных. Среди индикаторных и сопутствующих

переменных могут быть как случайные, так и не случайные (контролируемые в эксперименте).

Основные теоретические и прикладные разработки по КА относятся к *линейным* моделям. В частности, если анализируется схема из n наблюдений со скалярным результирующим признаком y, с k возможными типами условий эксперимента и с p сопутствующими переменными $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$, то линейная модель соответствующего КА задается уравнениями:

$$y_i = (\theta_{д1} \cdot x_{дi}^{(1)} + \ldots + \theta_{дk} \cdot x_{дi}^{(k)}) + (\theta_1 (X_{дi}) \cdot x_i^{(1)} +$$

$$+ \ldots + \theta_p (X_{дi}) \cdot x_i^{(p)}) + \varepsilon_i (X_{дi}), \quad i = 1, 2, \ldots, n, \qquad (13.28)$$

где индикаторные переменные $x_{дi}^{(j)} = 1$, если j-е условие эксперимента имело место при i-м наблюдении, и равны нулю — в противном случае; коэффициенты $\theta_{дj}$ определяют эффект влияния j-го условия; $x_i^{(s)}$ — значение сопутствующей переменной $x^{(s)}$, при котором наблюдался результирующий признак y_i ($i = 1, 2, \ldots, n$; $s = 1, 2, \ldots, p$); $\theta_s (X_{дi})$ — значения соответствующих коэффициентов регрессии y по $x^{(s)}$, вообще говоря, зависящие от конкретного сочетания условий эксперимента, т. е. от вектора $X_{дi} = (x_{дi}^{(1)}, \ldots, x_{дi}^{(k)})'$, а $\varepsilon_i(X_{дi})$ — величина остаточных случайных компонент («ошибок измерения»), имеющих нулевые средние значения. Основное содержание КА — в построении статистических оценок для неизвестных параметров $\theta_{д1}, \ldots, \theta_{дk}$; $\theta_1, \ldots, \theta_p$ и статистических критериев, предназначенных для проверки различных гипотез относительно значений этих параметров.

Если в (13.28) постулировать априори $\theta_1, = \ldots = \theta_p \equiv \equiv 0$, то получится модель дисперсионного анализа; если же из (13.28) исключить влияние неколичественных факторов (т. е. положить $\theta_{д1} = \ldots = \theta_{дk} = 0$), то получится линейная модель регрессионного анализа. Своим названием КА обязан тому обстоятельству, что в его вычислениях используются разбиения *ковариаций* переменных y и X точно так же, как в дисперсионном анализе используются разбиения остаточной суммы квадратов.

Считается, что термин «КА» введен Р. А. Фишером в связи с рассмотрением одной частной схемы этой модели в § 49 144-го издания книги «Статистические методы для исследователей» (пер. с англ. — М.: Статистика, 1958).

Весьма полные сведения по современным методам КА можно найти в [29, 66, 119, 148].

13.5.2. Оценивание неизвестных значений параметров и проверка гипотез в модели КА. Запишем линейную модель КА (13.28) в матричном виде:

$$Y = \mathbf{X}_д \Theta_д + \mathbf{X}\Theta + \boldsymbol{\varepsilon},$$

или[1]

$$Y = (\mathbf{X}_д, \mathbf{X}) \begin{pmatrix} \Theta_д \\ \Theta \end{pmatrix} + \boldsymbol{\varepsilon}, \tag{13.28'}$$

где Y — $(n \times 1)$-вектор-столбец наблюдений результирующего показателя; $\mathbf{X}_д$ — $(n \times k)$-матрица плана эксперимента по неколичественным факторам $X_д$; $\Theta_д$ — $(k \times 1)$-вектор-столбец неизвестных параметров, соответствующих неколичественным факторам (общее среднее, главные эффекты, взаимодействия и т. п.); $\mathbf{X}$ — $(n \times p)$-матрица плана регрессионных (количественных) объясняющих переменных; Θ — $(p \times 1)$-вектор-столбец параметров (неизвестных коэффициентов регрессии); $\boldsymbol{\varepsilon}$ — $(n \times 1)$-вектор-столбец случайных остатков модели, подчиняющийся нормальному распределению $N(\mathbf{0}, \sigma^2 \times \mathbf{I}_n)$, где остаточная дисперсия σ^2 неизвестна (подлежит оцениванию). Предполагается, что тип условий эксперимента $\mathbf{X}_д$ («способ обработки» — в исходной терминологии ДА) не влияет на матрицу плана регрессионных экспериментов $\mathbf{X}$, т. е. столбцы матрицы $\mathbf{X}$ линейно не зависят от столбцов матрицы $\mathbf{X}_д$ (*существенное* предположение). К *несущественным* предположениям относятся допущения о том, что матрицы $\mathbf{X}_д$ и $\mathbf{X}$ имеют полный ранг (соответственно k и p) и что не имеется ограничений на параметры $\Theta_д$ (о простых модификациях описываемых процедур в случае отказа от этих допущений см., например, [119, п. 3.8.3]).

Для нахождения оценок $\widehat{\Theta}_д$ и $\widehat{\Theta}$ неизвестных параметров $\Theta_д$ и Θ можно было бы формально рассмотреть (13.28′) как одну большую модель регрессии и применить к ней обычный

---

[1] Запись $(\mathbf{A}, \mathbf{B})$, где $\mathbf{A}$ и $\mathbf{B}$ — матрицы с одинаковым количеством строк, означает матрицу, полученную присоединением столбцов матрицы $\mathbf{B}$ к столбцам матрицы $\mathbf{A}$. Аналогично $\begin{pmatrix} \mathbf{C} \\ \mathbf{D} \end{pmatrix}$ — это матрица, полученная присоединением к строкам матрицы $\mathbf{C}$ строк матрицы $\mathbf{D}$ (где $\mathbf{C}$ и $\mathbf{D}$ — матрицы с одинаковым количеством столбцов). Существенное отличие моделей (13.28)—(13.28′) от внешне похожих на них моделей регрессионного и *классического* ковариационного анализа — в зависимости коэффициентов $\boldsymbol{\Theta}$ от неколичественных переменных $X_д$. В этом случае анализ моделей (13.28) осуществляется с помощью специальных методов расщепления смесей.

метод наименьших квадратов (см. § 7.1, гл. 11, а также [14, п. 8.6.3]). Однако можно добиться существенного упрощения анализа за счет использования специального строения матрицы ($\mathbf{X}_д$, $\mathbf{X}$) и наших знаний специфики модели ДА. С этой целью используется так называемый *двухшаговый метод наименьших квадратов* (подробнее о нем см. гл. 14). Этот метод (применительно к модели КА) состоит из следующих этапов:

1. В модели (13.28′) полагаем $\Theta \equiv 0$ и находим по описанным в § 13.2—13.4 правилам оценки $\widehat{\Theta_д}(0)$ и остаточную сумму квадратов (при условии $\Theta \equiv 0$):

$$\text{ОСК}(0) = Y'\mathbf{Q}Y, \tag{13.29}$$

где $\mathbf{Q} = \mathbf{I}_n - \mathbf{X}_д(\mathbf{X}_д'\mathbf{X}_д)^{-1}\mathbf{X}_д'$.

2. Заменяем в (13.29) Y на $Y - \mathbf{X}\cdot\Theta$ и находим такое $\widehat{\Theta}$, которое минимизирует полученное выражение. Итак,

$$\text{ОСК}(\Theta) = (Y - \mathbf{X}\Theta)'\,\mathbf{Q}\,(Y - \mathbf{X}\Theta);$$

$$\frac{\partial\{\text{ОСК}(\Theta)\}}{\partial\Theta} = 2\mathbf{X}'\mathbf{Q}\mathbf{X}\cdot\Theta - 2\mathbf{X}'\mathbf{Q}Y = \mathbf{0},$$

откуда

$$\widehat{\Theta} = (\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\cdot\mathbf{X}'\mathbf{Q}Y. \tag{13.30}$$

3. Подсчитывается остаточная сумма квадратов для *общей* модели (13.28′) ковариационного анализа, равная [119, п. 3.7.1]:

$$\text{ОСК} = \min_{\Theta}\text{ОСК}(\Theta) = (Y - \mathbf{X}\cdot\widehat{\Theta})'\,\mathbf{Q}\,(Y - \mathbf{X}\widehat{\Theta}) =$$

$$= Y'\mathbf{Q}Y - \widehat{\Theta}'\mathbf{X}'\mathbf{Q}Y. \tag{13.31}$$

4. Для получения оценок $\widehat{\Theta}_д$ заменяем в выражении для $\widehat{\Theta}_д(0)$ вектор Y вектором $Y - \mathbf{X}\cdot\widehat{\Theta}$.

Проверка гипотез относительно параметров $\theta_{дi}$ проводится так же, как в моделях ДА, только со значением ОСК, подсчитанным по формуле (13.31) и с числом степеней свободы k, равным числу степеней свободы ОСК модели ДА минус ранг матрицы $\mathbf{X}$. Проверка гипотезы $H_\Theta : \Theta \equiv \mathbf{0}$ проводится с помощью статистики

$$\left(\frac{1}{\text{ранг}(\mathbf{X})}\cdot\widehat{\Theta}'\mathbf{X}'\mathbf{Q}Y\right)\Big/\frac{1}{k}\cdot\text{ОСК},$$

которая в предположении справедливости гипотезы H_Θ имеет $F(t, k)$-распределение (t-ранг $(\mathbf{X})$).

13.5.3. Связь с проблемой статистического исследования смесей многомерных распределений. Посмотрим на модель регрессии результирующего показателя η по объясняющим переменным $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ как на одну из характеристик их закона распределения, например, функции плотности

$$p\,(y;\, x^{(1)}, \ldots, x^{(p)};\, \Theta\,(X_{д}),\, \mathbf{V}\,(X_{д})), \qquad (13.32)$$

зависящей от параметров регрессии Θ и ковариационной матрицы «остатков» $\mathbf{V}$, которые в свою очередь *зависят от типа условий эксперимента* $X_{д}$[1]. Тогда, анализируя данные вида

$$\{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)};\, y_i\}_{i=\overline{1,n}},$$

зафиксированные *при различных* типах условий эксперимента $X_{дi}$, в действительности имеем дело с выборкой *из смеси распределений* вида (13.32), поскольку при варьировании типа условий $X_{дi}$ меняются и значения параметров Θ и $\mathbf{V}$, от которых зависит анализируемый закон распределения, а следовательно, меняется и вид искомой регрессионной зависимости $f\,(X) = \mathbf{E}\,(\eta\,|X)$.

Игнорирование этого обстоятельства является причиной многих недоразумений и неудач в прикладных исследованиях, опирающихся на аппарат регрессионного анализа. Для объяснения этого обстоятельства представим себе, что при исследовании линейной парной регрессионной зависимости исходные данные $\{(x_i, y_i)\}_{i=\overline{1,n}}$ фиксировались при *переключающемся* (в неизвестные для исследователя моменты времени) режиме типа условий эксперимента: либо в режиме 1, в котором (при весьма высокой корреляции) регрессия имела *монотонно возрастающий* характер, либо в режиме 2, в котором (при столь же высокой корреляции) регрессия имела монотонно убывающий характер (см. рис. 13.1). Очевидно, попытки выявить связь между y и x по такой *смешанной* выборке не увенчаются успехом: вычисления покажут, что связи нет. В то же время, если предварительно (или одновременно с решением задач регрессии) разбить имеющиеся данные на однородные (по условиям эксперимента) подвыборки и *строить функции регрес-*

---

[1]Если объясняющие переменные случайны, то речь идет о $(p+1)$-мерной плотности, зависящей от параметров Θ и $\mathbf{V}$. Если $x^{(1)}, \ldots, x^{(p)}$ неслучайны, то они интерпретируются как варьируемые (но известные) значения параметров, от которых зависит одномерная плотность $p\,(y;\, X;\, \Theta,\, \mathbf{V})$.

сии отдельно для каждой такой подвыборки, то удастся установить тесную статистическую зависимость между исследуемыми переменными.

Ковариационный анализ предоставляет исследователю один из возможных подходов к реализации описанной схемы. Другие подходы опираются на *статистический анализ смесей многомерных распределений*: оценку параметров смеси распределений [11], модели *типологической регрессии* [4, 11, 82]. Подробное описание этих методов предполагается дать в следующем томе данного издания.

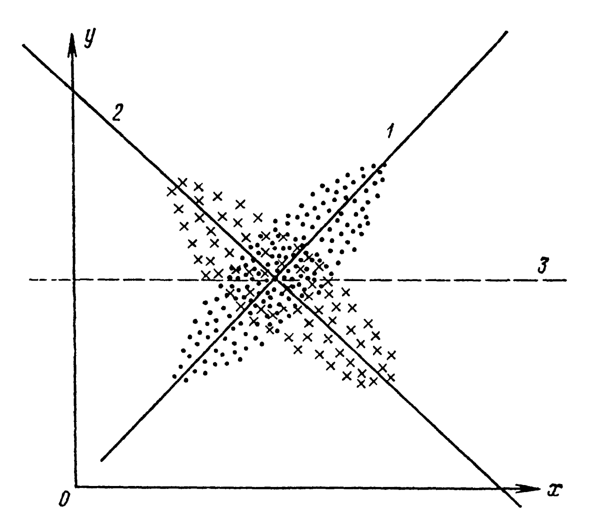


Рис. 13.1. Прямые 1, 2 и 3 — графики аппроксимирующих функций регрессии, построенных соответственно по наблюдениям подвыборок: 1 (точки), 2 (крестикн) и по объединенной выборке, состоящей из тех и других наблюдений

13.6. Влияние нарушений основных предположений

Мы уже видели ранее, что отклонения от предположения нормальности распределения могут существенно сказываться на эффективности оценок среднего и дисперсии [14, п. 8.6.1 и 10.4.4]. Проиллюстрируем еще раз этот факт на примере влияния эксцесса на величину доверительного интервала для s^2,

построенного по выборке большого объема. Согласно [14, п. 8.6.5] в нормальном случае доверительный интервал строится по величине $(n-1)s^2/\sigma^2$, имеющей распределение $\chi^2(n-1)$. Для s^2/σ^2 в случае нормального распределения

$$\mathbf{E}(s^2/\sigma^2) = 1; \quad (13.33)$$

$$\mathbf{D}(s^2/\sigma^2) = 2/(n-1). \quad (13.34)$$

При отличном от нуля эксцессе β_2 [14, п. 5.6.6] при любом n ситуация сильно меняется, так как, хотя (13.33) и остается тем же, вместо (13.34) имеем:

$$\mathbf{D}(s^2/\sigma^2) = 2/(n-1) + \beta_2/n. \quad (13.35)$$

При больших n отношение (13.34) к (13.35) приближается к $1 + \beta_2/2$, и поскольку s^2 имеет приблизительно нормальное распределение, легко может быть подсчитан доверительный интервал для σ^2. В табл. 13.1 для разных значений β_2 показана вероятность P того, что истинное значение σ^2 не попадает в 95%-ный доверительный интервал, построенный согласно нормальной теории при больших значениях n.

Таблица 13.5 [148]

Эксцесс β_2	$-1,5$	-1	$-0,5$	0	0,5	1	2	4	7
P	$9\cdot10^{-5}$	$6\cdot10^{-3}$	0,024	0,05	0,08	0,11	0,17	0,26	0,36

Как видно по таблице, для реально встречающихся на практике распределений (см. [14, п. 6.1.11]) истинная ошибка первого рода может быть очень большой, в несколько раз превышая нормальную ошибку в 5%.

Будем называть выводы, относящиеся только к постоянным факторам ДА, выводами о средних, а выводы, относящиеся к случайным эффектам, — выводами о дисперсиях. Примером первых являются критерии для проверки гипотез о главных эффектах и взаимодействиях в моделях с постоянными факторами и соответствующие доверительные интервалы. Примером выводов о дисперсиях являются критерии равенства дисперсий в моделях со случайными факторами и доверительные интервалы для компонент дисперсии. Нарушение предпо-

ложений нормальности оказывает слабое влияние на выводы о средних и очень опасно при выводах о дисперсиях. Первым на это обратил внимание Е. Пирсон [148, 234].

Другая опасность, подстерегающая исследователя при использовании ДА, — это не отраженная в модели коррелированность между наблюдениями. Рассмотрим простейшую модель корреляции между последовательными наблюдениями. Предполагается, что cor $(x_i, x_{i+1}) = \rho$ для $i = 1, \ldots, n - 1$, а все остальные коэффициенты корреляции равны нулю. Возможны все ρ, такие, что $|\rho| \leqslant 0,5$. Некоторым обоснованием этого предположения является наблюдение Стьюдента [148, § 10.1], вычислившего коэффициенты корреляции между последовательными анализами пяти различных химических свойств с выборками из одной и той же партии хорошо перемешанного материала: 0,27; 0,31; 0,19; 0,09; 0,09.

Для иллюстрации влияния отклонений от предположения независимости воспользуемся тем же методическим приемом, что и выше, а именно: построим при большом n доверительный интервал для среднего μ исходя из предположения независимости наблюдений и подсчитаем P — вероятность того, что 95%-ный доверительный интервал не накроет истинное значение μ. Результаты показаны в табл. 13.2. Из нее следует один вывод: неучтенная корреляция последовательных наблюдений может серьезно влиять на статистические выводы.

Таблица 13.6 [148]

ρ	−0,4	−0,3	−0,2	−0,1	0	0,1	0,2	0,3	0,4
P	10^{-5}	0,002	0,011	0,028	0,050	0,074	0,098	0,12	0,14

Значительная неучтенная корреляция может возникнуть в моделях со случайными факторами при иерархической классификации (§ 13.4), когда при построении модели и планировании сбора данных пропускается один из источников варьирования результатов экспериментов.

Влияние неравенства дисперсий наблюдений изучалось многими авторами. Общий вывод [148, § 10.4]: в моделях с постоянными факторами его надо учитывать только в случае плохо сбалансированного распределения экспериментальных точек.

ВЫВОДЫ

1. *Дисперсионным анализом* называется метод организации (планирования), статистического анализа и интерпретации результатов экспериментов, в которых изучается зависимость количественной переменной y от сочетания градаций качественных переменных $X_д$. В ДА используются линейные модели с постоянными и случайными факторами. Дисперсионный анализ с постоянными факторами можно рассматривать как специальный случай регрессионного анализа. Свое название ДА получил из-за того, что при проверке гипотез о влиянии на y изучаемых факторов используется разложение суммы квадратов $\sum_i (y_i - \bar{y})^2$ на слагаемые, соответствующие проверяемым гипотезам.

2. Простейшая модель ДА с одним постоянным фактором имеет вид $y_{ij} = \theta_0 + \theta_i + \varepsilon_{ij}$, $j = 1, ..., J_i$; $i = 1, ..., I$, где $\sum_i \theta_i = 0$, $\varepsilon_{ij} \in N(0, \sigma^2)$ и независимы между собой. Основная гипотеза H: $\theta_1 = ... = \theta_I = 0$. Для ее проверки используется критерий $F = (I-1)^{-1} \cdot \sum_i J_i \cdot (y_{i*} - y_{**})^2 / [(n - I)^{-1} \sum_i \sum_j (y_{ij} - y_{**})^2]$, где $n = \sum_i J_i$; $y_{i*} = \sum_j y_{ij}/J_i$, $y_{**} = \sum_i \sum_j y_{ij}/n$. Если H верна, то F имеет $F(I-1, n-I)$-распределение. Если H отвергается, то изучают, насколько θ_i отличаются друг от друга. Для этого строят *одновременные* доверительные интервалы для сумм $\sum_i c_i\theta_i$ $(\sum_i c_i = 0)$, называемых *сравнениями*.

3. Основными схемами организации ДА с двумя факторами являются: 1) *перекрестная классификация*, в которой каждая градация одного фактора сочетается в эксперименте с каждой градацией другого с математической моделью вида $\mathrm{E}y_{ij} = \theta + \alpha_i + \beta_j + \gamma_{ij}$, где $\sum_i \alpha_i = \sum_j \beta_j = \sum_j \gamma_{ij} = 0$; 2) *иерархическая* (*гнездовая*) классификация с моделью вида $\mathrm{E}y_{ij} = \theta + \alpha_i + \beta_{ij}$, где $\sum_i \alpha_i = 0$, $\sum_j \beta_{ij} = 0$. Большинство используемых на практике моделей ДА может быть получено путем различных сочетаний указанных основных схем организаций экспериментов.

4. Схемы ДА с постоянными факторами довольно устойчивы к нарушениям предположений о распределении случайных ошибок (нормальность, равенство дисперсий). Этого нельзя сказать о схемах ДА со случайными факторами. Нарушения

(фактически непроверяемых) предположений о характере распределения эффектов случайных факторов (нормальность, независимость) могут привести к существенным ошибкам при проверке гипотез и построении доверительных интервалов.

5. *Ковариационным анализом* (КА) называется совокупность методов организации (планирования), статистического анализа и интерпретации результатов эксперимента или наблюдений, в которых изучается зависимость количественной переменной y от сочетания градаций (типов условий эксперимента) качественных переменных $X_д$ и одновременно от набора количественных объясняющих переменных X, которые в данной схеме называются сопутствующими.

6. Основной метод, используемый при оценивании неизвестных параметров модели КА, — это *двухшаговый метод наименьших квадратов.*

7. Если качественные переменные $X_д$ в схеме КА неконтролируемы, то задача сводится к исследованию моделей регрессионного анализа на основании выборки, извлеченной из *смеси генеральных совокупностей* (количество компонент смеси равно числу типов условий эксперимента, при которых регистрировались выборочные данные). Эта задача предусматривает предварительное (или одновременное с процессом построения искомых регрессий) разбиение исходной выборки на однородные (по условиям эксперимента) части и оценку функции регрессии *отдельно для каждой такой части.*

Раздел IV. СИСТЕМЫ ОДНОВРЕМЕННЫХ УРАВНЕНИЙ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ АППАРАТА СТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ ЗАВИСИМОСТЕЙ

Глава 14. ОЦЕНИВАНИЕ ПАРАМЕТРОВ СИСТЕМ ОДНОВРЕМЕННЫХ ЭКОНОМЕТРИЧЕСКИХ УРАВНЕНИЙ

14.1. Системы одновременных уравнений

14.1.1. Определение и специфика проблематики систем одновременных уравнений. При изучении функционирования экономических систем исследователь обычно сталкивается со следующей ситуацией.

Состояние системы в каждый момент времени t описывается набором переменных $Z_{t_1}, \dots, Z_{t_M}$, среди которых есть как *эндогенные* (внутрисистемные), так и *экзогенные* (внешние по отношению к рассматриваемой системе). Между переменными существуют функциональные и статистические связи. К первому типу относятся тождества, вытекающие из определений и содержательного смысла переменных. Ко второму типу относятся поведенческие связи, являющиеся выражениями экономических законов, действующих в системе. Поскольку поведение экономических систем носит статистический характер (присутствуют случайные возмущения, погрешности, неучтенные факторы), то для описания поведенческих связей используются регрессионные уравнения. В теории экономико-статистического моделирования систему взаимосвязанных регрессионных уравнений и тождеств, в которой одни и те же переменные в различных регрессионных уравнениях могут одновременно выступать и в роли результирующих показателей, и в роли объясняющих переменных, принято называть системой одновременных (эконометрических) уравнений. При этом в соотношения могут входить переменные, относящиеся не только к периоду t, но и к предшествующим периодам, называемые лаговыми («запаздывающими») переменными.

Для экономистов большой интерес представляет количественный анализ модели, т. е. нахождение оценок параметров на основании имеющейся в распоряжении исследователя информации о значениях переменных. Первая из возникающих здесь проблем: можно ли в предложенной модели однозначно

восстановить значение некоторого параметра или же его определение принципиально невозможно на основе рассматриваемой модели? Это так называемая проблема идентифицируемости[1] — первоочередная на этапе формирования модели, поскольку прежде, чем переходить к процедурам оценивания, необходимо быть уверенным, что их применение имеет смысл.

Проблема оценивания здесь также имеет свои особенности. Основная трудность состоит в том, что в эконометрических моделях переменная, играющая роль независимой (объясняющей) переменной в одном соотношении, может быть зависимой в другом. Это приводит к тому, что в регрессионных уравнениях системы объясняющие переменные и случайные возмущения оказываются, вообще говоря, коррелированными. Наконец, в современной практике встречаются модели, имеющие десятки и даже сотни уравнений (в том числе и нелинейных), в связи с чем возникают и вычислительные трудности.

Указанные обстоятельства обусловили необходимость построения специальной теории, изучающей статистический аспект экономико-математических моделей. К настоящему времени довольно хорошо разработан ее раздел, относящийся к моделям, описываемым системами линейных уравнений. Имеется ряд итоговых монографий (содержащих обширную библиографию), среди которых отметим [46, 80, 106, 143]. Гл. 13 и 14 [46], посвященные проблемам идентифицируемости и оценивания для систем одновременных уравнений, могут быть рекомендованы для первоначального ознакомления с предметом.

Современные исследования в указанной области составляют содержание журналов «Econometrica» и «Journal of Econometrics».

14.1.2. Два традиционных примера. Прежде чем перейти к формулировке общей линейной модели, рассмотрим в качестве иллюстраций два классических примера.

Пример 14.1. Рассмотрим модель спроса и предложения («крест Маршалла»). Спрос Q на некоторый продукт и его предложение Q' зависят от цены p продукта. Рыночный механизм формирует цену таким образом, что спрос и предложение уравниваются. Наблюдению доступна равновесная цена и спрос (совпадающий с предложением).

Возникающая здесь линейная модель выглядит следующим образом:

$$Q_t = \alpha_1 p_t + \beta_1 + u_t$$

[1]В литературе распространен также термин «проблема идентификации», который представляется менее удачным, поскольку зачастую «идентификация» используется как синоним термина «оценивание».

(«спрос пропорционален цене»), $\alpha_1 < 0$;

$\tilde{Q}_t = \alpha_2 p_t + \beta_2 + u_t'$

(«предложение пропорционально цене»), $\alpha_2 > 0$;

$Q_t = \tilde{Q}_t$.

Величины u_t, u_t', $t = 1, \ldots, n$, — случайные возмущения, имеющие нулевые средние. Оказывается, без дополнительных предположений (например, на структуру случайных возмущений) интересующие экономистов параметры α_i и β_i однозначно в этой модели определить нельзя, т. е. они *неидентифицируемы*. Точные определения и утверждения даны в следующем параграфе.

Пример 14.2. Содержательный смысл модели спроса состоит в утверждении, что потребительские расходы, т. е. спрос, пропорциональны доходу. В свою очередь доход есть сумма потребительских и непотребительских расходов.

Математическая формулировка модели такова:

$$c_t = \alpha + \beta y_t + u_t; \qquad (14.1)$$

$$y_t = c_t + z_t, \qquad (14.2)$$

где c — потребительские расходы; y — доход; z — непотребительские расходы; u — случайное возмущение (учитывающее неполноту информации, незамкнутость системы и т. п.). Предполагается, что уровень непотребительских расходов задан извне, т. е. переменная z экзогенна и определяется независимо от c и y. Случайные величины u_t, $t = 1, \ldots, n$, некоррелированы, имеют нулевые средние и одинаковые дисперсии σ^2. Требуется оценить параметры модели α, β, σ^2.

В (14.1) переменная y коррелирует со случайным возмущением. В самом деле,

$$y_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} z_t + \frac{u_t}{1-\beta}.$$

Поэтому

$$\mathbf{E} u_t (y_t - \mathbf{E} y_t) = \frac{1}{1-\beta} \mathbf{E} u_t^2 \neq 0,$$

и посылки метода наименьших квадратов для уравнения (14.1) не выполняются. Это приводит к тому, что обычные мнк-оценки параметров (14.1) оказываются смещенными и несостоятельными. Например, мнк-оценка параметра β имеет вид

$$\hat{\beta} = \frac{\beta m_{zz} + (1+\beta) m_{zu} + m_{uu}}{m_{zz} + 2m_{zu} + m_{uu}}, \qquad (14.3)$$

где

$$m_{zz} = \frac{1}{n} \sum_{t=1}^{n} z_t^2, \quad m_{zu} = \frac{1}{n} \sum_{t=1}^{n} z_t u_t, \quad m_{uu} = \frac{1}{n} \sum_{t=1}^{n} u_t^2 .$$

Предполагая, как обычно, что $P\text{-}\lim m_{zu} = 0$[1] (это соотношение всегда справедливо в силу закона больших чисел, если z_t — детерминированные), $P\text{-}\lim m_{zz} = \bar{m}_{zz}$, где $\bar{m}_{zz}$ — некоторая константа, имеем

$$P\text{-}\lim \hat{\beta} = \beta + (1-\beta) \frac{\sigma^2/\bar{m}_{zz}}{1+\sigma^2/\bar{m}_{zz}}$$

и, значит, $P\text{-}\lim \hat{\beta} > \beta$, если $\beta < 1$, и $P\text{-}\lim \hat{\beta} < \beta$, если $\beta > 1$.

Таким образом, при $\beta \neq 1$ оценка (14.3) несостоятельна.

14.1.3. Общая линейная модель. В данной главе мы будем изучать линейную модель вида

$$\beta_{i1} y_{1t} + \beta_{i2} y_{2t} + \ldots + \beta_{iG} y_{Gt} + \gamma_{i1} x_{1t} + \ldots + \gamma_{iK} x_{Kt} = u_{it}; \qquad (14.4)$$

$t = 1, \ldots, n; \; i = 1, \ldots, G.$

Здесь y_{it} — значения эндогенных переменных в момент t; x_{it} — значения экзогенных переменных в момент t и лаговых эндогенных переменных. Переменные x_{it} в момент времени t называются *предопределенными*.

Совокупность равенств (14.4) называется *системой одновременных уравнений в структурной форме*. На коэффициенты (14.4) накладываются априорные ограничения, например, часть коэффициентов считаются равными нулю. Это и обеспечивает возможность статистическиого оценивания оставшихся.

Систему (14.4) удобно переписать в следующем матричном виде:

$$\mathbf{B} y_t + \mathbf{\Gamma} x_t = u_i, \qquad (14.5)$$

где $\mathbf{B}$ — матрица порядка $G \times G$, состоящая из коэффициентов при текущих значениях эндогенных переменных; $\mathbf{\Gamma}$ — матрица порядка $G \times K$, состоящая из коэффициентов при предопределенных переменных;

$$y_t = (y_{1t}, \ldots, y_{Gt})'; \; x_t = (x_{1t}, \ldots, x_{Kt})'; \; u_t = (u_{1t}, \ldots, u_{Gt})'$$

— вектор-столбцы.

---

[1] $P\text{-}\lim$ — обозначение для предела по вероятности.

Предположим, что матрица **B** невырождена. Тогда уравнение (14.5) можно разрешить относительно y_t:

$$y_t = \mathbf{\Pi}\, x_t + \eta_t, \tag{14.6}$$

где $\mathbf{\Pi} = -\mathbf{B}^{-1}\mathbf{\Gamma}$ — матрица размерности $G \times K$; $\eta_t = \mathbf{B}^{-1} u_t$ — случайное возмущение. Равенство (14.6) называется *приведенной формой* системы одновременных уравнений.

Введем обозначения:

$$\mathbf{Y} = \begin{bmatrix} y_1' \\ \vdots \\ y_n' \end{bmatrix}; \mathbf{X} = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}; \mathbf{U} = \begin{bmatrix} u_1' \\ \\ u_n' \end{bmatrix}. \tag{14.7}$$

Тогда все уравнения (14.5) для всех периодов наблюдений могут быть записаны в виде одного матричного уравнения

$$\mathbf{Y}\mathbf{B}' + \mathbf{X}\,\mathbf{\Gamma}' = \mathbf{U}. \tag{14.8}$$

Эта компактная форма будет использоваться в дальнейшем.

14.2. Спецификация модели и проблема идентифицируемости

14.2.1. Идентифицируемость приведенной формы. Мы будем рассматривать модель, описываемую системой одновременных уравнений, имеющих структурную форму вида

$$\mathbf{B}\, y_t + \mathbf{\Gamma}\, x_t = u_t, \quad t = 1, \dots, n, \tag{14.9}$$

где матрица **B** порядка $G \times G$ невырождена. Для простоты будем считать, вообще говоря, стохастические переменные x_t экзогенными.

Спецификация модели помимо списка эндогенных и экзогенных переменных включает в себя априорную информацию: ограничения на коэффициенты и гипотезу о случайных возмущениях u_t, а также правило нормализации.

Типичным примером априорных ограничений являются исключающие ограничения, выражающие то, что некоторые переменные заведомо не входят в отдельные уравнения и, следовательно, соответствующие им коэффициенты равны нулю.

В качестве гипотезы о случайных возмущениях примем, что случайные величины u_t независимы и имеют один и тот же закон распределения μ_u с нулевым средним.

Предположим, что i-е уравнение может быть разрешено относительно y_{it} и масштабы коэффициентов выбраны так, что

коэффициент при y_{it} в i-м уравнении равен единице. Это и есть правило нормализации.

В рамках данной спецификации фиксируем какую-либо структуру S, т. е. матрицы $\mathbf{B}$, $\mathbf{\Gamma}$ и распределение μ_u случайного возмущения (параметрами структуры будут элементы матриц $\mathbf{B}$, $\mathbf{\Gamma}$ и само распределение μ_u). Тогда вектор y_t при заданном x_t будет иметь некоторое распределение $P^S(x_t)$. Поскольку далее x_t фиксировано, мы будем употреблять для этого распределения более короткое обозначение P_t^S.

Структуры S и $\tilde{S}$ называются (наблюдаемо) эквивалентными, если $P_t^S = P_t^{\tilde{S}}$, $t = 1, \ldots, n$.

Параметр α называется идентифицируемым в структуре S, если $\alpha = \tilde{\alpha}$ для любой структуры $\tilde{S}$, эквивалентной S. Иными словами, параметр α идентифицируем, если из равенств $P_t^S = P_t^{\tilde{S}}$, $t = 1, \ldots, n$, следует, что $\alpha = \tilde{\alpha}$.

Структура S называется идентифицируемой, если все ее параметры идентифицируемы.

Рассмотрим вопрос об идентифицируемости приведенной формы

$$y_t = \mathbf{\Pi}\, x_t + \eta_t, \tag{14.10}$$

связанной со структурной формой соотношениями

$$\mathbf{\Pi} = -\mathbf{B}^{-1}\,\mathbf{\Gamma}, \quad \eta_t = \mathbf{B}^{-1}\, u_t.$$

П р е д л о ж е н и е 1. Пусть $\operatorname{rank} \mathbf{X} = K$ — числу экзогенных переменных. Тогда структура приведенной формы (14.10) является идентифицируемой.

Доказательство. Пусть $S = (\mathbf{\Pi}, \mu_\eta)$ и $\tilde{S} = (\tilde{\mathbf{\Pi}}, \mu_{\tilde{\eta}})$ — две эквивалентные структуры, т. е. $P_t^S = P_t^{\tilde{S}}$, $t \leqslant n$. Наряду с (14.10) имеем, что $\tilde{y}_t = \tilde{\mathbf{\Pi}} x_t + \tilde{\eta}_t$, где случайные векторы $\tilde{\eta}_t$ подчиняются распределению $\mu_{\tilde{\eta}}$. По предположению $\mathbf{E} y_t = \mathbf{E} \tilde{y}_t$. Поскольку η_t и $\tilde{\eta}_t$ имеют нулевые математические ожидания, то $\mathbf{\Pi} x_t = \tilde{\mathbf{\Pi}} x_t$. Введем матрицу $\mathbf{X}_K$, столбцами которой являются K линейно независимых векторов $x_{t_1}, \ldots, x_{t_K}$. Тогда $(\mathbf{\Pi} - \tilde{\mathbf{\Pi}})\,\mathbf{X}_K = \mathbf{0}$ и, значит, в силу невырожденности $\mathbf{X}_K$, $\mathbf{\Pi} = \tilde{\mathbf{\Pi}}$.

Наконец, имеет место и равенство распределений $\mu_\eta = \mu_{\tilde{\eta}}$, поскольку $\eta_t = y_t - \mathbf{\Pi} x_t$, $\tilde{\eta}_t = \tilde{y}_t - \mathbf{\Pi} x_t$, а распределения P_t^S и $P_t^{\tilde{S}}$ случайных векторов y_t и $\tilde{y}_t$ совпадают.

14.2.2. Проблема идентифицируемости для структурной формы. Пусть $S = (\mathbf{A}, \mu_u)$, $\tilde{S} = (\tilde{\mathbf{A}}, \mu_{\tilde{u}})$, где $\mathbf{A} = [\mathbf{B}, \mathbf{\Gamma}]$, $\tilde{\mathbf{A}} = [\tilde{\mathbf{B}}\tilde{\mathbf{\Gamma}}]$.

Предложение 2. Пусть $\operatorname{rank} \mathbf{X} = K$. Структуры S и $\tilde{S}$ эквивалентны тогда и только тогда, когда существует невырожденная $G \times G$ матрица $\mathbf{D}$, такая, что $\mathbf{A} = \mathbf{D}\tilde{\mathbf{A}}$ и распределение u_t совпадает с распределением $\mathbf{D}\tilde{u}_t$ (т. е. $\mu_{\tilde{u}} = \mu_u \mathbf{D}^{-1}$).

Доказательство. Пусть матрица $\mathbf{D}$ с указанными свойствами существует; $y_t = \mathbf{\Pi} x_t + \eta_t$, $\tilde{y}_t = \tilde{\mathbf{\Pi}} x_t + \eta_t$ — приведенные формы, отвечающие структурам S и $\tilde{S}$. Поскольку $\mathbf{B} = \mathbf{D}\tilde{\mathbf{B}}$ и $\mathbf{\Gamma} = \mathbf{D}\tilde{\mathbf{\Gamma}}$, то

$$\mathbf{\Pi} = -\mathbf{B}^{-1}\mathbf{\Gamma} = \tilde{\mathbf{B}}^{-1}\mathbf{D}^{-1}\mathbf{D}\tilde{\mathbf{\Gamma}} = \tilde{\mathbf{B}}^{-1}\tilde{\mathbf{\Gamma}} = \tilde{\mathbf{\Pi}}.$$

Распределения случайных векторов η и $\tilde{\eta}$ также совпадают:

$$\mu_\eta = \mu_u \mathbf{B} = \mu_u \mathbf{D}\tilde{\mathbf{B}} = \mu_{\tilde{u}} \mathbf{D}^{-1}\mathbf{D}\tilde{\mathbf{B}} = \mu_{\tilde{u}} \mathbf{B} = \mu_{\tilde{\eta}}.$$

Следовательно, совпадают и распределения векторов y_t и $\tilde{y}_t$.

Обратно, пусть S и $\tilde{S}$ — эквивалентные структуры. Положим $\mathbf{D} = \mathbf{B}\tilde{\mathbf{B}}^{-1}$. Тогда $\mathbf{B} = \mathbf{D}\tilde{\mathbf{B}}$. Согласно предложению 1 из совпадения распределений y_t и $\tilde{y}_t$ при всех $t \leqslant n$ вытекает совпадение матриц $\mathbf{\Pi}$ и $\tilde{\mathbf{\Pi}}$ и распределений μ_η и $\mu_{\tilde{\eta}}$.
Но равенство $-\mathbf{B}^{-1}\mathbf{\Gamma} = -\tilde{\mathbf{B}}^{-1}\tilde{\mathbf{\Gamma}}$ влечет равенство $\mathbf{\Gamma} = \mathbf{D}\tilde{\mathbf{\Gamma}}$. В свою очередь соотношение $\mu_\eta = \mu_{\tilde{\eta}}$ означает, что $\mu_u \mathbf{B} = \mu_{\tilde{u}} \mathbf{B}$, откуда вытекает требуемое свойство: $\mu_u = \mu_{\tilde{u}} \mathbf{D}^{-1}$.
Далее мы будем предполагать выполненным условие предложений 1 и 2 об отсутствии мультиколлинеарности у экзогенных переменных: $\operatorname{rank} \mathbf{X} = K$.

Предложение 1 показывает, что весь класс эквивалентных структур обладает одной и той же приведенной формой. Из этого следует, что коэффициент матрицы $\mathbf{A} = [\mathbf{B}\mathbf{\Gamma}]$ будет идентифицируемым, если априорные ограничения обеспечивают однозначность его восстановления по матрице приведенной формы.

Для решения проблемы идентифицируемости можно воспользоваться и предложением 2, из которого следует, что множество матриц структурных коэффициентов во всех структурах, эквивалентных данной структуре S, получается умножением $\mathbf{A}$ слева на невырожденные матрицы из некоторого клас-

са M, который определяется априорными ограничениями. Интересующий нас коэффициент будет идентифицируем тогда и только тогда, когда он инвариантен относительно преобразований структурной формы матрицами из M.

14.2.3. Критерии идентифицируемости. Рассмотрим некоторые наиболее важные типы ограничений и приведем критерии идентифицируемости, используемые в практических задачах.

Будем предполагать, что априорные ограничения являются линейными однородными функциями, каждая из которых зависит только от коэффициентов одного из уравнений структурной формы. Выясним, когда коэффициенты матрицы **А** могут быть однозначно восстановлены по матрице приведенной формы **П**.

Пусть $\mathbf{I}_K$ — единичная матрица порядка $K \times K$. Введем обозначение

$$\mathbf{W} = \begin{bmatrix} \mathbf{\Pi} \\ \mathbf{I}_K \end{bmatrix}.$$

Соотношение $\mathbf{B\Pi} + \mathbf{\Gamma} = \mathbf{0}$ между структурной и приведенной формой теперь может быть записано в виде

$$\mathbf{AW} = \mathbf{0}, \qquad (14.11)$$

где **А** — матрица порядка $G \times (G + K)$; **W** — матрица порядка $(G + K) \times K$, имеющая ранг K.

Пусть α_1 — первая строка матрицы **А**. Тогда из (14.11) следует, что

$$\alpha_1 \mathbf{W} = 0. \qquad (14.12)$$

Равенство (14.12) представляет собой систему из K (независимых) уравнений относительно $G + K$ неизвестных — элементов вектора α_1.

Согласно предположению априорные ограничения на элементы могут быть записаны в виде

$$\alpha_1 \mathbf{\Phi} = 0, \qquad (14.13)$$

где $\mathbf{\Phi}$ — матрица из $G + K$ строк, имеющая столько столбцов, сколько ограничений.

Например, пусть априори известно, что коэффициент $\beta_{12} = 0$. Тогда один из столбцов матрицы $\mathbf{\Phi}$ имеет вид $(0, 1, 0, \ldots, 0)'$.

Из (14.12) и (14.13) следует, что элементы вектора α_1 являются решениями системы уравнений

$$\alpha_1 [\mathbf{W\Phi}] = 0. \qquad (14.14)$$

В силу правила нормализации ($\beta_{11} = 1$) для идентифицируемости первого уравнения необходимо и достаточно, чтобы пространство решений системы (14.14) было одномерным, т. е.

$$\operatorname{rank}[\mathbf{W\Phi}] = G + K - 1. \tag{14.15}$$

Пусть L — число ограничений. Тогда $[\mathbf{W\Phi}]$ содержит $K + L$ столбцов и для выполнения (14.15) необходимо, чтобы $L \geqslant G - 1$:

для идентифицируемости какого-либо из уравнений необходимо, чтобы число ограничений было не меньше числа уравнений модели, уменьшенного на единицу.

Если имеются только исключающие ограничения, т. е. априорная информация о равенстве нулю некоторых коэффициентов, то необходимое условие идентифицируемости определенного уравнения таково:

число неизвестных, исключенных из уравнения, должно быть по меньшей мере равно числу уравнений минус единица. Последнее условие может быть сформулировано следующим образом:

число исключенных из уравнения экзогенных переменных должно быть не меньше числа участвующих в нем эндогенных переменных, уменьшенного на единицу.

Сформулированные необходимые условия (так называемые правила порядка) в силу своей простоты являются весьма полезными при решении проблемы идентифицируемости, поскольку при построении модели они позволяют сразу выявить неидентифицируемые уравнения. Однако эти условия могут оказаться далекими от достаточных. Необходимое и достаточное условие (14.15) не годится для проверки идентифицируемости модели, поскольку требует построения матрицы $\mathbf{\Pi}$. Тем не менее из него можно извлечь критерий идентифицируемости и в терминах структурной формы (правило ранга).

Первое уравнение системы идентифицируемо тогда и только тогда, когда $\operatorname{rank}(\mathbf{A\Phi}) = G - 1$. Это утверждение может быть выведено непосредственно из (14.15), но мы получим его из соображений, связанных с инвариантностью коэффициентов при умножении структурной формы на допустимые матрицы.

Запишем (14.9) в виде

$$\mathbf{A}z_t = u_t, \tag{14.16}$$

где $z_t = [y_t' x_t']'$.

Применяя к (14.16) невырожденную $(G \times G)$-матрицу $\mathbf{D}$, получим преобразованную структурную форму $\mathbf{DA}z_t = \mathbf{D}u_t$. Матрица $\mathbf{D}$ допустима, если преобразованная структурная

форма удовлетворяет априорным ограничениям, которые могут быть записаны в виде $\alpha_1 \mathbf{\Phi} = 0$ или, эквивалентно, $e_1 (\mathbf{A\Phi}) = 0$,

где $e_1 = (1, 0, \ldots, 0)$ — вектор длины G. Первая строка преобразованной структурной формы имеет вид $d_1 \mathbf{A}$, где d_1 — первая строка матрицы $\mathbf{D}$. Поэтому

$$d_1 (\mathbf{A\Phi}) = 0. \tag{14.17}$$

Идентифицируемость первого уравнения структурной формы означает, что $e_1 \mathbf{A} = d_1 \mathbf{A}$. В частности, $e_1 \mathbf{B} = d_1 \mathbf{B}$. Так как матрица невырождена, то $e_1 = d_1$. Таким образом (с учетом правила нормализации), вектор-строка d_1 из (14.17) определяется однозначно; с точностью до пропорциональности решение (14.17) есть e_1. Следовательно, $\operatorname{rank} (\mathbf{A\Phi}) = G - 1$. В свою очередь это равенство влечет идентифицируемость первого уравнения. Проиллюстрируем полученные выводы примерами.

Пусть имеется следующая система, состоящая из двух уравнений:

$$\beta_{11} y_{1t} + \beta_{12} y_{2t} + \gamma_{11} x_{1t} + \gamma_{12} x_{2t} = u_{1t};$$

$$\beta_{21} y_{1t} + \beta_{22} y_{2t} + \gamma_{21} x_{1t} + \gamma_{22} x_{2t} = u_{2t}.$$

В силу соглашения о правиле нормализации $\beta_{11} = \beta_{22} = 1$.

Пример 14.3. Предположим, что априорные ограничения касаются только коэффициентов матрицы $\mathbf{\Gamma}: \gamma_{12} = 0$, $\gamma_{22} = 0$. Отметим, что именно эти ограничения имеются в модели спроса и предложения (пример 14.1).

Для первого уравнения системы $\mathbf{\Phi} = (0, 0, 0, 1)'$, $\mathbf{A\Phi} = (0, 0)'$. Ранг $\mathbf{A\Phi}$ равен 0 и, следовательно, первое уравнение идентифицируемо (здесь $G - 1 = 1 \neq 0$). То же самое имеет место и для второго уравнения.

Пример 14.4. Пусть $\gamma_{12} = 0$, $\gamma_{21} = 0$.

Для первого уравнения $\mathbf{\Phi} = (0, 0, 0, 1)'$ и $\mathbf{A\Phi} = (\gamma_{12}, \gamma_{22})' = (0, \gamma_{22})'$. При условии, что $\gamma_{22} \neq 0$, мы имеем $\operatorname{rank} (\mathbf{A\Phi}) = 1 = G - 1$. Следовательно, первое уравнение идентифицируемо; аналогичный вывод можно сделать и для другого уравнения.

Решая уравнение $\alpha_1 [\mathbf{W\Phi}] = 0$, которое в данном случае имеет вид

$$(\beta_{11}, \beta_{12}, \gamma_{11}, \gamma_{12}) \begin{bmatrix} \pi_{11} & \pi_{12} & 0 \\ \pi_{21} & \pi_{22} & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \mathbf{0},$$

и, пользуясь равенством $\beta_{11} = 1$, получаем однозначное выражение параметров первого уравнения через параметры приведенной формы.

Пример 14.5. Пусть $\gamma_{11} = 0$, $\gamma_{12} = 0$.
Для первого уравнения

$$\mathbf{\Phi} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \mathbf{A\Phi} = \begin{bmatrix} 0 & 0 \\ \gamma_{21} & \gamma_{22} \end{bmatrix}.$$

Следовательно, rank $(\mathbf{A\Phi}) = 1$ (в предположении $\gamma_{21}\gamma_{22} \neq 0$), и первое уравнение идентифицируемо. Записывая подробно уравнение $\alpha_1 [\mathbf{W\Phi}] = 0$, получаем систему

$$\begin{aligned} \beta_{11}\pi_{11} + \beta_{12}\pi_{21} + \gamma_{11} &= 0; \\ \beta_{11}\pi_{12} + \beta_{12}\pi_{22} = \gamma_{12} &= 0; \\ \gamma_{11} &= 0; \\ \gamma_{12} &= 0, \end{aligned}$$

откуда

$$\beta_{11} = 1, \ \beta_{12} = -\pi_{11}/\pi_{21} = -\pi_{12}/\pi_{22}.$$

Таким образом, для того, чтобы не возникло противоречие, должно выполняться определенное соотношение между коэффициентами приведенной формы. Встретившаяся ситуация, когда имеются ограничения на коэффициенты приведенной формы, носит название *сверхидентифицируемости*.

Если мы будем оценивать коэффициенты матрицы приведенной формы без учета ограничений, а затем из полученных оценок образуем оценку для β_{12} (т. е. применим так называемый косвенный метод наименьших квадратов), то выражения $-\widehat{\pi}_{11}/\widehat{\pi}_{21}$ и $-\widehat{\pi}_{12}/\widehat{\pi}_{22}$ не будут равны друг другу. Косвенный метод наименьших квадратов непригоден для оценивания в случае сверхидентифицируемости.

При рассмотрении проблемы идентифицируемости мы ограничились случаем, когда имеются априорные ограничения только на структурные коэффициенты. Ясно, что ограничения на вид распределения случайных возмущений могут сузить класс M допустимых преобразований так, что неидентифицируемое без этих ограничений уравнение станет идентифицируемым.

14.3. Рекурсивные системы

Среди систем одновременных уравнений наиболее простыми являются рекурсивные системы, для оценивания коэффициентов которых можно применять обыкновенный метод наименьших квадратов.

Система одновременных уравнений

$$\mathbf{B}y_t + \mathbf{\Gamma}x_t = u_t \tag{14.18}$$

называется рекурсивной, если выполнены следующие условия:

1) матрица $\mathbf{B}$ является нижней треугольной матрицей, т. е. $\beta_{ij} = 0$ при $j > i$, $\beta_{ii} = 1$;

2) ковариационная матрица возмущений диагональна:

$\mathbf{E}\,uu' = \mathbf{\Sigma} = (\sigma_{ij})$, $\sigma_{ij} = 0$ при $i \neq j$;

3) каждое ограничение на структурные коэффициенты относится к отдельному уравнению.

Предложение 3. Пусть $\operatorname{rank} \mathbf{X} = K$ и матрица $\mathbf{\Sigma}$ невырождена, т. е. $\sigma_{ii} > 0$ при $i = 1, \ldots, G$. Тогда структурная форма рекурсивной системы идентифицируема.

Доказательство. Пусть $S = (\mathbf{B}, \mathbf{\Gamma}, \mu_u)$, $\tilde{S} = (\mathbf{B}, \mathbf{\Gamma}, \mu_{\tilde{u}})$ — две эквивалентные структуры. Согласно предложению 2 существует матрица $\mathbf{D}$, такая, что

$$\mathbf{B} = \mathbf{D}\tilde{\mathbf{B}},\ \mathbf{\Sigma} = \mathbf{D}\,\mathbf{\Sigma}\,\mathbf{D}'. \tag{14.19}$$

Матрица $\tilde{\mathbf{B}}^{-1}$ представляет собой нижнюю треугольную матрицу с единицами на главной диагонали. Такой же будет и матрица $\mathbf{D} = \mathbf{B}\tilde{\mathbf{B}}^{-1}$, т. е. $d_{ij} = 0$, $j > i$, $d_{ii} = 1$. Из второго равенства в (14.19) вытекает, что $d_{ij}\sigma_{jj} = 0$, $i \neq j$, следовательно, $d_{ij} = 0$ и когда $i > j$.

Значит, $\mathbf{D}$ — единичная матрица; структуры S и $\tilde{S}$ совпадают, что и требовалось доказать.

Покажем, что процедура оценивания коэффициентов структурной формы рекурсивной системы методом наименьших квадратов, примененным к отдельному уравнению, приводит к состоятельным оценкам.

Запишем i-е уравнения рекурсивной системы для всех n периодов наблюдений в следующем виде:

$$y_{(i)} = \mathbf{Z}\,\theta + u_{(i)}, \tag{14.20}$$

где $y_{(i)} = (y_{i1}, \ldots, y_{in})'$, $\mathbf{Z} = [\mathbf{Y}_{(i)}\mathbf{X}]$; $\mathbf{Y}_{(i)}$ — матрица, состоящая из первых $i - 1$ столбцов матрицы $\mathbf{Y}$, которая вместе с $\mathbf{X}$ была введена в (14.7), $u_{(i)} = (u_{i1}, \ldots, u_{in})'$. Оцениванию

подлежит вектор коэффициентов $\theta = -(\beta_{i1}, \ldots, \beta_{i,i-1}, \gamma_{i1}, \ldots, \gamma_{iK})'$.

Будем предполагать, что выполнены следующие условия

а) $P\text{-}\lim_{n} \frac{1}{n} \mathbf{X}' \mathbf{U} = 0$;

в) $P\text{-}\lim_{n} \frac{1}{n} \mathbf{Z}' \mathbf{Z} = \Sigma_{zz}$,

где Σ_{zz} — невырожденная матрица.

Предложение 4. При выполнении условий а) и в) мнк-оценка параметра θ для (14.20) состоятельна.

Доказательство. Прежде всего покажем, что

$$P\text{-}\lim \frac{1}{n} \mathbf{Y}'_{(i)} u_{(i)} = 0. \tag{14.21}$$

Транспонируем равенство (14.8), умножим его затем слева на $\mathbf{B}^{-1}$, а справа на $\mathbf{U}/n$. Переходя к пределу по вероятности при $n \to \infty$ и пользуясь условием а), получаем в результате, что

$$P\text{-}\lim \frac{1}{n} \mathbf{Y}' \mathbf{U} = \mathbf{B}^{-1} \Sigma = \Sigma_{yu}. \tag{14.22}$$

Матрица Σ диагональна, $\mathbf{B}^{-1}$ — нижняя треугольная матрица, поэтому Σ_{yu} — также нижняя треугольная матрица. Следовательно, равенство (14.22) влечет (14.21).

Соотношение (14.21) показывает, что случайное возмущение в i-м уравнении не коррелирует в пределе с эндогенными переменными, входящими в это уравнение. Поэтому для i-го уравнения с точки зрения оценивания переменные $y_1, \ldots, y_{i-1}$ ничем не отличаются от предопределенных переменных, что и приводит к состоятельности оценки $\widehat{\theta}$ наименьших квадратов, которая здесь имеет вид

$$\widehat{\theta} = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' y_{(i)} = \theta + (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' u_{(i)}.$$

Ясно, что $P\text{-}\lim_{n} \widehat{\theta} = \theta$, поскольку в силу а), в) и (14.21)

$$P\text{-}\lim_{n} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' u_{(i)} = P\text{-}\lim_{n} \left(\frac{1}{n} \mathbf{Z}' \mathbf{Z}\right)^{-1} =$$

$$= P\text{-}\lim_{n} \frac{1}{n} \mathbf{Z}' u_{(i)} = \Sigma_{zz}^{-1} \cdot 0 = 0.$$

Доказательство завершено.

Указанные выше привлекательные свойства рекурсивных систем вызывают желание использовать именно их в эконометрических исследованиях. К тому же можно привести аргументы в пользу того, что реальные экономические системы являются рекурсивными по своей природе.

Например, многие экономисты считают модель примера 14.1 несовершенной. Действительно, трудно представить себе рынок, где равновесные цены и спрос формировались бы одновременно. Реалистичнее выглядит следующая ситуация. Цены в день t устанавливаются в зависимости от объема продаж в предыдущий день, в то время как покупки в день t зависят от цены товара в день t.

Математическая формализация приводит к системе

$$p_t = \alpha_0 + \alpha_1 q_{t-1} + u_t;$$

$$q_t = \beta_0 + \beta_1 p_t + v_t,$$

где случайные возмущения u_t и v_t можно считать независимыми, т. е. к рекурсивной системе.

Необходимость рассматривать системы, отличные от рекурсивных, возникает в связи с тем, что исследователь обычно располагает некоторыми усредненными (агрегированными) данными. Например, данные о рыночной конъюнктуре могут быть усреднены по недельным или месячным периодам. Предположим, что известны величины: P_t — средняя цена за неделю t и Q_t — средний объем ежедневных продаж за неделю t. Если считать время реакции рынка, как и раньше, равным одному дню, то соотношение $P_t = \alpha_0 + \alpha_1 Q_{t-1} + u_t$ вряд ли можно признать разумным. В этой ситуации рассмотренная в § 14.1 модель представляется более естественной.

14.4. Двух- и трехшаговый методы наименьших квадратов

14.4.1. Наиболее распространенные методы оценивания системы одновременных уравнений. Формальное применение мнк для получения оценок коэффициентов системы одновременных уравнений приводит, вообще говоря, к оценкам с плохими статистическими свойствами — смещенным и несостоятельным. Поэтому область его применения ограничена рекурсивными системами. Для оценивания параметров точно идентифицируемой системы можно применить косвенный метод наименьших квадратов, состоящий в оценивании обычным мнк коэффициентов приведенной формы и подстановке оценок в выра-

жения для коэффициентов структурной формы через коэффициенты приведенной формы, что приводит к смещенным, но состоятельным оценкам. В случае сверхидентифицируемости косвенный метод наименьших квадратов, как отмечалось в 14.2, не применим.

Для оценивания произвольных систем одновременных уравнений в настоящее время имеется довольно значительное количество методов, которые делятся на две группы. К первой группе относятся методы, применимые к каждому уравнению в. отдельности: *двухшаговый метод наименьших квадратов* (2 мнк), метод максимума правдоподобия с ограниченной информацией, называемый также методом наименьшего дисперсионного соотношения [46] или методом Комиссии Коулса [80], и некоторые другие. Вторая группа содержит методы, предназначенные для оценивания всей системы в целом. Это методы максимума правдоподобия и *трехшаговый метод наименьших квадратов* (3 мнк). Несколько особняком стоят итеративные методы, или методы неподвижной точки, которые обладают определенными вычислительными достоинствами, что немаловажно при исследовании систем большой размерности, однако статистические их свойства изучены в недостаточной степени.

14.4.2. Двухшаговый метод наименьших квадратов. Наиболее важным методом оценивания отдельного уравнения системы, получившим широкое распространение, является двухшаговый метод наименьших квадратов. Он дает состоятельные, вообще говоря, смещенные оценки коэффициентов, является достаточно простым с теоретической точки зрения и удобен для вычислений.

Запишем интересующее нас уравнение (для определенности, первое системы (14.4)) в виде

$$y = \mathbf{Y}_1 \beta + \mathbf{X}_1 \gamma + u, \tag{14.23}$$

где $y = (y_{1t}, \ldots, y_{1n})'$ — вектор n наблюдений над переменной y_1; $\mathbf{Y}_1$ — матрица $n \times g$ наблюдений над другими текущими значениями эндогенных переменных, входящих в уравнение; β — вектор размерности $g \times 1$ структурных коэффициентов, относящихся к переменным из матрицы $\mathbf{Y}_1$; $\mathbf{X}_1$ — матрица порядка $n \times k$ наблюдений над предопределенными переменными, входящими в уравнение; γ — вектор размерности $k \times 1$ коэффициентов, относящихся к переменным из матрицы $\mathbf{X}_1$; u — вектор случайных возмущений, имеющий размерность $n \times 1$.

Рассмотрим часть приведенной формы, состоящую из уравнений, которые выражают эндогенные переменные, входящие

в правую часть (14.23), через предопределенные переменные системы:

$$\mathbf{Y}_1 = \mathbf{X}\mathbf{\Pi}_1' + \mathbf{V}_1. \qquad (14.24)$$

Пусть $\widehat{\mathbf{\Pi}}_1$ — мнк-оценка матрицы $\mathbf{\Pi}_1'$, т. е.

$$\widehat{\mathbf{\Pi}}_1' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1,$$

где $\mathbf{X} = [\mathbf{X}_1\mathbf{X}_2]$; $\mathbf{X}_2$ — матрица наблюдений над предопределенными переменными, не входящими в (14.23).

Положим

$$\widehat{\mathbf{Y}}_1 = \mathbf{X}\widehat{\mathbf{\Pi}}_1 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1; \qquad (14.25)$$

$$\widehat{\mathbf{V}}_1 = \mathbf{Y}_1 - \widehat{\mathbf{Y}}_1. \qquad (14.26)$$

Тогда уравнение (14.23) может быть записано в следующем виде:

$$y = \mathbf{Z}\delta + \varepsilon, \qquad (14.27)$$

где $\mathbf{Z} = [\widehat{\mathbf{Y}}_1\mathbf{X}_1]$, $\delta = [\beta'\gamma']'$, $\varepsilon = u + \widehat{\mathbf{V}}_1\beta$.

Применим к (14.27) метод наименьших квадратов. В результате для вектора неизвестных коэффициентов δ получаем оценку

$$d = \begin{bmatrix} b \\ c \end{bmatrix} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}y = \begin{bmatrix} \widehat{\mathbf{Y}}_1'\mathbf{Y}_1 & \widehat{\mathbf{Y}}_1'\mathbf{X}_1 \\ \mathbf{X}_1'\widehat{\mathbf{Y}}_1 & \mathbf{X}_1'\mathbf{X} \end{bmatrix}^{-1} \begin{bmatrix} \widehat{\mathbf{Y}}_1' y \\ \mathbf{X}_1' y \end{bmatrix},$$

которая в исходных переменных имеет вид

$$\begin{bmatrix} b \\ c \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_1 & \mathbf{Y}_1'\mathbf{X}_1 \\ \mathbf{X}_1'\mathbf{Y}_1 & \mathbf{X}_1'\mathbf{X}_1 \end{bmatrix}^{-1} \times$$

$$\times \begin{bmatrix} \mathbf{Y}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}y \\ \mathbf{X}_1' y \end{bmatrix}. \qquad (14.28)$$

Эта оценка и носит название оценки двухшагового метода наименьших квадратов параметров β и γ.

Таким образом, существо двухшагового метода состоит в замене матрицы $\mathbf{Y}_1$ расчетной матрицей $\widehat{\mathbf{Y}}_1$, после чего искомые коэффициенты вычисляются обыкновенной регрессией y на $\widehat{\mathbf{Y}}_1$ и $\mathbf{X}$.

Далее мы будем предполагать, что матрица $\mathbf{Y}_1$ состоит из наблюдений за эндогенными переменными с номерами 2, ..., $g+1$, матрица $\mathbf{X}_1$ — из наблюдений за предопределен-

ными переменными с номерами 1, ..., k (этого можно добиться изменением нумерации). Теперь введенные выше обозначения согласуются с обозначениями 14.1, 14.2, причем

$$\mathbf{\Pi}_1 = \begin{bmatrix} \pi_{21} & \dots & \pi_{2K} \\ \pi_{g+1,1} & \dots & \pi_{g+1,K} \end{bmatrix},$$

а первая строка матрицы $\mathbf{A} = [\mathbf{B\Gamma}]$ имеет вид

$$\alpha_1 = (1, -\beta_1, \dots, -\beta_g, 0, \dots, 0, -\gamma_1, \dots, -\gamma_k, 0, \dots 0)$$

(для уравнения (14.23) мы имеем только исключающие ограничения).

Пусть $\mathbf{I}_k$ — единичная матрица $k \times k$; $\mathbf{O}_k$ — нулевая матрица порядка $k \times (K-k)$; $\mathbf{\Theta} = [\mathbf{I}_k \mathbf{O}_k]$. Рассмотрим $(g+k) \times K$-матрицу

$$\tilde{\mathbf{\Pi}}_1 = \begin{bmatrix} \mathbf{\Pi}_1 \\ \mathbf{\Theta} \end{bmatrix}.$$

Лемма 1. Пусть уравнение (14.23) идентифицируемо. Тогда

$$\operatorname{rank} \tilde{\mathbf{\Pi}}_1 = g + k.$$

Доказательство. Из (14.14) вытекает, что α_1 является единственным решением системы уравнений

$$\alpha_1 \mathbf{\Delta} = e_1,$$

где $\mathbf{\Delta} = [e_1' \mathbf{W\Phi}]$, $e_1 = (1, 0, \dots, 0)$ — вектор длины $G + K$; $\mathbf{\Phi}$ — матрица, с помощью которой записаны соответствующие исключающие ограничения: $\alpha_{1,\,g+2} = 0, \dots, \alpha_{1,\,G} = 0$, $\alpha_{1,\,G+k+1} = 0, \dots, \alpha_{1,\,G+K} = 0$. Значит, $\operatorname{rank} \mathbf{\Delta} = G + K$, и все строки $\mathbf{\Delta}$ линейно независимы. Линейная независимость строк с номерами $2, \dots, g+1$, $G+1, \dots, G+k$ влечет утверждение леммы.

Предложение 5. Пусть уравнение (14.23) идентифицируемо и выполнены следующие условия:

а) $P\text{-}\lim\limits_{n} \frac{1}{n} \mathbf{X}' \mathbf{U} = \mathbf{0}$;

б) $P\text{-}\lim\limits_{n} \frac{1}{n} \mathbf{X}' \mathbf{X} = \Sigma_{xx}$,

где Σ_{xx} — невырожденная матрица.

Условие а) всегда выполнено в силу закона больших чисел, когда x_t детерминированы. Тогда 2 мнк-оценка состоятельна.

Доказательство. Из (14.25)—(14.27) следует, что

$$d = \delta + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(u + \widehat{\mathbf{V}}_1\beta) = \delta + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'u.$$

В силу предположений а) и в)

$$P\text{-}\lim_n \widehat{\mathbf{\Pi}}_1' = \mathbf{\Pi}_1', \tag{14.29}$$

поэтому

$$P\text{-}\lim_n \frac{1}{n}\widehat{\mathbf{Y}}_1' u = P\text{-}\lim_n \frac{1}{n}\widehat{\mathbf{\Pi}}_1\mathbf{X}_1' u = \mathbf{\Pi}_1 \cdot P\text{-}\lim_n \frac{1}{n}\mathbf{X}_1' u = 0.$$

Следовательно,

$$P\text{-}\lim_n \frac{1}{n}\mathbf{Z}'u = P\text{-}\lim_n \frac{1}{n}\begin{bmatrix}\widehat{\mathbf{Y}}_1' u\\ \mathbf{X}_1' u\end{bmatrix} = 0,$$

и нам осталось показать только, что $P\text{-}\lim \frac{1}{n}\mathbf{Z}'\mathbf{Z}$ является невырожденной матрицей.

Имеем в силу условия б) и соотношения (14.29), что

$$\mathbf{\Sigma}_{zz} = P\text{-}\lim_n \frac{1}{n}\mathbf{Z}'\mathbf{Z} = P\text{-}\lim_n \begin{bmatrix}\widehat{\mathbf{\Pi}}_1\\ \mathbf{\Theta}\end{bmatrix}\frac{1}{n}\mathbf{X}'\mathbf{X}\,[\widehat{\mathbf{\Pi}}_1'\ \mathbf{\Theta}'] =$$

$$= \widetilde{\mathbf{\Pi}}_1 \mathbf{\Sigma}_{xx} \widetilde{\mathbf{\Pi}}_1',$$

и невырожденность $\mathbf{\Sigma}_{zz}$ вытекает из леммы 1 и известного свойства положительно определенных матриц.

14.4.3. Трехшаговый метод наименьших квадратов. Этот метод применяется для оценки параметров системы одновременных уравнений в целом. Сначала к каждому уравнению применяется двухшаговый метод с целью оценить коэффициенты и возмущения каждого структурного уравнения, а затем построить оценку для ковариационной матрицы возмущений. После этого для оценивания коэффициентов всей системы применяется обобщенный метод наименьших квадратов.

Приведем формальное описание трехшагового мнк. Для этого рассмотрим систему одновременных уравнений, содержащую G эндогенных и K предопределенных переменных, которые будем считать нестохастическими. Запишем i-е уравнение в виде

$$y_i = \mathbf{Y}_i\beta_i + \mathbf{X}_i\gamma_i + u_i, \tag{14.30}$$

где y_i — вектор, образованный из $n+1$ наблюдений над зависимой переменной с номером i; $\mathbf{Y}_i$ — матрица порядка $n \times g_i$, составленная из наблюдений за остальными эндоген-

ными переменными, входящими в это уравнение; $\mathbf{X}_i$ — матрица порядка $n \times k_i$ наблюдений над предопределенными переменными i-го уравнения.

В обозначениях

$$\mathbf{Z}_i = [\mathbf{Y}_i\ \mathbf{X}_i], \qquad \delta_i = \begin{bmatrix} \beta_i \\ \gamma_i \end{bmatrix}$$

i-е уравнение системы переписывается следующим образом:

$$y_i = \mathbf{Z}_i \delta_i + u_i. \tag{14.31}$$

Если $\mathbf{X}$ — матрица порядка $n \times K$, образованная из наблюдений за всеми предопределенными переменными, то

$$\mathbf{X}' y_i = \mathbf{X}' \mathbf{Z}_i \delta_i + v_i. \tag{14.32}$$

Ковариационная матрица возмущений $v_i = \mathbf{X}' u_i$ имеет вид

$$\mathbf{E} v_i v_i' = \mathbf{E} \mathbf{X}' u_i u_i' \mathbf{X} = \sigma_{ii} \mathbf{X}' \mathbf{X},$$

где σ_{ii} — дисперсия случайного возмущения, испытываемого i-м уравнением исходной системы.

Применяя к (14.32) обобщенный метод наименьших квадратов, получаем оценку

$$d_i = [\mathbf{Z}_i' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}_i]^{-1} \mathbf{Z}_i' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' y_i$$

вектора коэффициентов δ_i, которая, как нетрудно убедиться, совпадает с 2 мнк-оценкой для (14.30).

Запишем все уравнения (14.32) в следующей форме:

$$\begin{bmatrix} \mathbf{X}' y_1 \\ \mathbf{X}' y_2 \\ \vdots \\ \mathbf{X}' y_G \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}' \mathbf{Z}_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \mathbf{X}' \mathbf{Z}_G \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_G \end{bmatrix} + \begin{bmatrix} \mathbf{X}' u_1 \\ \mathbf{X}' u_2 \\ \vdots \\ \mathbf{X}' u_G \end{bmatrix}. \tag{14.33}$$

Если бы матрица ковариаций Σ была известна, то формальное применение обобщенного метода наименьших квадратов

к (14.33) привело бы к оценке параметров системы, содержащей матрицу ковариаций вектора возмущений

$$\mathbf{V}=\begin{bmatrix} \sigma_{11}\mathbf{X}'\mathbf{X} & \sigma_{12}\mathbf{X}'\mathbf{X}\ldots & \sigma_{1G}\mathbf{X}'\mathbf{X} \\ \sigma_{21}\mathbf{X}'\mathbf{X} & \sigma_{22}\mathbf{X}'\mathbf{X}\ldots & \sigma_{2G}\mathbf{X}'\mathbf{X} \\ \vdots & \vdots & \vdots \\ \sigma_{G1}\mathbf{X}'\mathbf{X} & \sigma_{G2}\mathbf{X}'\mathbf{X}\ldots & \sigma_{GG}\mathbf{X}'\mathbf{X} \end{bmatrix}=\Sigma\otimes\mathbf{X}'\mathbf{X}$$

с $\mathbf{V}^{-1}=\Sigma^{-1}\otimes(\mathbf{X}'\mathbf{X})^{-1}$.

Идея трехшагового метода наименьших квадратов состоит в использовании вместо Σ ее оценки $\mathbf{S}=(s_{ij})$, где

$$s_{ij}=\frac{(y_i-\mathbf{Z}_i d_i)'(y_i-\mathbf{Z}_j d_j)}{(n-k_i-g_i)^{1/2}(n-k_j-g_j)^{1/2}}.$$

В результате приходим к искомой 3 мнк-оценке

$$\hat{\delta}=\left\{\begin{bmatrix} \mathbf{Z}_1'\mathbf{X} & 0 & \ldots & 0 \\ 0 & \mathbf{Z}_2'\mathbf{X} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \mathbf{Z}_G'\mathbf{X} \end{bmatrix}\times\right.$$

$$\left.\times\,\Sigma^{-1}\otimes(\mathbf{X}'\mathbf{X})^{-1}\begin{bmatrix} \mathbf{X}'\mathbf{Z}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{X}'\mathbf{Z}_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & 0 \\ 0 & 0 & \ldots & \mathbf{X}'\mathbf{Z}_G \end{bmatrix}\right\}^{-1}\times$$

$$\times\begin{bmatrix} \mathbf{X}'\mathbf{Z}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{X}'\mathbf{Z}_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \mathbf{X}'\mathbf{Z}_G \end{bmatrix}\Sigma^{-1}\otimes(\mathbf{X}'\mathbf{X})^{-1}\begin{bmatrix} \mathbf{X}'y_1 \\ \mathbf{X}'y_2 \\ \vdots \\ \mathbf{X}'y_G \end{bmatrix},$$

которая после упрощений может быть записана в виде

$$\hat{\delta}=\begin{bmatrix} s^{11}\mathbf{Z}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_1 \ldots & s^{1G}\mathbf{Z}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_1 \\ \vdots & \vdots \\ s^{G1}\mathbf{Z}_G'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_G \ldots & s^{GG}\mathbf{Z}_G'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_G \end{bmatrix}\times$$

$$\times\begin{bmatrix} \sum_{j=1}^{G} s^{ij}\mathbf{Z}_1'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y_j \\ \sum_{j=1}^{G} s^{Gj}\mathbf{Z}_j'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y_j \end{bmatrix},$$

где s^{ij} — элементы матрицы $\mathbf{S}^{-1}$.

В случае, когда матрица Σ не является диагональной, т. е. когда одновременные возмущения, входящие в различные структурные уравнения, зависимы, трехшаговая процедура имеет лучшую асимптотическую эффективность по сравнению с двухшаговой.

14.5. Метод неподвижной точки

Для оценивания параметров систем одновременных уравнений в настоящее время помимо классических методов, рассмотренных в предыдущих параграфах, используются различные итеративные процедуры, основанные на методе неподвижной точки.

Запишем общую линейную модель (14.7), (14.8) в следующем виде:

$$\mathbf{Y} = \mathbf{YC} + \mathbf{XD} + \mathbf{U},$$

где матрица $\mathbf{C}$ получена из $-\mathbf{B}'$ заменой диагональных элементов последней нулями, $\mathbf{D} = -\mathbf{\Gamma}'$. Соответствующая приведенная форма имеет вид

$$\mathbf{Y} = \mathbf{XP} + \mathbf{V}$$

с

$$\mathbf{P} = \mathbf{D}(\mathbf{I} - \mathbf{C})^{-1}, \quad \mathbf{V} = \mathbf{U}(\mathbf{I} - \mathbf{C})^{-1}.$$

Положим $\mathbf{Y}^* = \mathbf{XD}(\mathbf{I} - \mathbf{C})^{-1}$. Тогда, очевидно,

$$\mathbf{Y} = \mathbf{Y}^*\mathbf{C} + \mathbf{XD} + \mathbf{V}; \qquad (14.34)$$

$$\mathbf{Y}^* = \mathbf{Y}^*\mathbf{C} + \mathbf{XD}. \qquad (14.35)$$

Соотношения (14.34), (14.35) называют переформулированной формой системы одновременных уравнений.

В методе неподвижной точки соотношение (14.34) рассматривается как регрессионное уравнение, в котором помимо неизвестных $\mathbf{C}$ и $\mathbf{D}$ неизвестными являются и объясняющие переменные $\mathbf{Y}^*$, для которых должно выполняться соотношение (14.35). Оказывается, при выполнении некоторых дополнительных условий $\mathbf{C}$, $\mathbf{D}$ и $\mathbf{Y}^*$ определяются однозначно (доказательство использует теорему о неподвижной точке, отсюда и название метода). Более того, они могут быть найдены при помощи следующей итеративной процедуры.

Начальное значение $\mathbf{Y}^0$ (нулевое приближение для $\mathbf{Y}^*$) берется произвольным, после чего первые приближения $\mathbf{C}^1$ и

$\mathbf{D}^1$ для $\mathbf{C}$ и $\mathbf{D}$ вычисляются из регрессии $\mathbf{Y}$ на $\mathbf{Y}^0$ и $\mathbf{X}$, т. е. из соотношения

$$\mathbf{Y}=\mathbf{Y}^0\mathbf{C}^1+\mathbf{X}\mathbf{D}^1+\mathbf{V}^1.$$

Следующее приближение $\mathbf{Y}^1$ для $\mathbf{Y}^*$ определяется равенством

$$\mathbf{Y}^1=\mathbf{Y}^0\mathbf{C}^1+\mathbf{X}\mathbf{D}^1.$$

Далее указанный процесс повторяется:

. . . .

$$\mathbf{Y}=\mathbf{Y}^{s-1}\mathbf{C}^s+\mathbf{X}\mathbf{D}^s+\mathbf{V}^s,$$

$$\mathbf{Y}^s=\mathbf{Y}^{s-1}\mathbf{C}^s+\mathbf{X}\mathbf{D}^s,$$

$$s=2, 3,\ldots$$

При этом

$$\mathbf{Y}^*=P\text{-}\lim_s \mathbf{Y}^s,\quad \mathbf{C}=P\text{-}\lim_s \mathbf{C}^s,\quad \mathbf{D}=P\text{-}\lim_s \mathbf{D}^s.$$

Таким образом, алгоритм метода неподвижной точки представляет из себя комбинацию метода наименьших квадратов и итерационного метода Якоби решения системы линейных алгебраических уравнений. На практике, однако, было установлено, что этот алгоритм далеко не всегда является сходящимся. Для улучшения сходимости были предложены различные его модификации. Например, в релаксационном методе неподвижной точки приближение $\mathbf{Y}^s$ для $\mathbf{Y}^*$ находится по формуле

$$\mathbf{Y}^s=\alpha\overline{\mathbf{Y}}^s+(1-\alpha)\mathbf{Y}^{s-1},$$

где $\overline{\mathbf{Y}}^s=\mathbf{Y}^{s-1}\mathbf{C}^s+\mathbf{X}\mathbf{D}^s$; α — некоторая постоянная, $\alpha\in$ $\in(0, 2)$. Имеются также модификации, цель которых состоит в уменьшении объема вычислений (рекурсивные методы неподвижной точки). Отметим, что теоретические свойства оценок, полученных по методу неподвижной точки, изучены недостаточно. Показано, однако, что если соответствующие итеративные процедуры сходятся, то полученные оценки параметров являются состоятельными. Исследованы свойства метода для малых выборок и произведено его сравнение с другими методами при помощи численного моделирования. Имеющиеся здесь результаты носят достаточно противоречивый и неполный характер. В экспериментах был замечен один большой недостаток метода [106]: оценки являются неустойчивыми и не всегда сходятся к какому-либо одному решению.

14.6. Сравнение методов

К настоящему времени предложено значительное количество методов состоятельного оценивания параметров систем одновременных эконометрических уравнений. Как отмечено в 14.4.1, они могут быть разбиты на две группы.

Первую группу составляют методы ограниченной информации. Представителями оценок этой группы являются 2 мнк-оценки (см. 14.4.2) и оценки максимального правдоподобия с ограниченной информацией. Можно показать, что 2 мнк-оценки и оценки максимального правдоподобия с ограниченной информацией асимптотически эквивалентны.

Вторую группу составляют методы, использующие полную информацию о системе, т. е. о строении ее уравнений и о степени их стохастической зависимости. Наиболее известными представителями этой группы являются трехшаговый метод наименьших квадратов, рассмотренный в 14.4.3, и метод максимального правдоподобия. Между оценками, получаемыми при помощи этих методов, существует тесная взаимосвязь: 3 мнк-оценки можно рассматривать в качестве первого приближения оценок метода максимума правдоподобия, по определению минимизирующих функцию плотности распределения наблюдений (в предположении, что они распределены по нормальному закону). Более того, указанные оценки асимптотически эквивалентны.

Основываясь на асимптотике оценок, можно утверждать, что если отклонения в уравнениях нормально распределены, то оценки с полной информацией для выборок большого объема будут более эффективны, чем оценки с ограниченной информацией. Однако уместна ли апелляция к асимптотическим свойствам оценок в условиях малых выборок и невозможности эффективной проверки гипотезы о нормальном распределении отклонений?

Другой важный вопрос связан с устойчивостью оценок по отношению к ошибкам спецификации, т. е. к неправильно выбранной форме связи, автокоррелированности или гетероскедастичности отклонений, нарушениям гипотезы о нормальности возмущений и т. д.

К сожалению, ответов, основанных на теоретических работах, на поставленные вопросы на сегодняшний день не существует. По-видимому, единственным инструментом исследования свойств оценок параметров одновременных эконометрических уравнений в условиях конечных выборок является метод статистических испытаний, или метод Монте-Карло.

Приведем некоторые общие выводы, на которых сходятся авторы большинства статистических экспериментов.

1. Все методы дают смещенные оценки.

2. При большом количестве наблюдений теоретические выводы, основанные на асимптотических свойствах, в основном подтверждаются.

3. В условиях малых выборок предпочтительность одного метода другому может меняться при переходе от одной эконометрической модели к другой. Иногда, вопреки всем статистическим аргументам, лучшим оказывается обыкновенный метод наименьших квадратов.

4. Оценки с ограниченной информацией оказываются более устойчивыми к ошибкам спецификации модели. Наоборот, оценки с полной информацией весьма чувствительно реагируют на изменения структуры.

На практике при оценивании параметров конкретных одновременных уравнений принимаются во внимание не только статистические свойства оценок, но и соображения вычислительного характера.

К сожалению, вплоть до настоящего времени одним из наиболее распространенных методов оценивания является обычный мнк (который, как показано в 14.1.2, не является состоятельным). Из состоятельных методов наиболее часто используется 2 мнк, который удобен и с точки зрения объема вычислений. Применение методов с полной информацией сдерживается большой размерностью задач оценивания (число уравнений достигает иногда несколько сотен и даже тысяч), а также отсутствием удовлетворительных пакетов программ.

ВЫВОДЫ

1. Одновременные уравнения возникают при изучении сложных систем, поведение которых описывается совокупностью законов, связывающих характеристики системы. Статистическое моделирование таких систем осуществляется при помощи регрессионных уравнений. При этом переменные, являющиеся объясняемыми в одном уравнении, в других уравнениях могут играть роль объясняющих.

2. Восстановление коэффициентов системы одновременных уравнений возможно лишь при наличии определенной априорной информации, например, равенства нулю каких-то коэффициентов или функций от них. Первый этап исследования модели направлен на то, чтобы ответить на вопрос, достаточно ли этой априорной информации, для чего используются критерии идентифицируемости (правила порядка и ранга). В слу-

чае неидентифицируемости обычно модель должна быть модифицирована.

3. Простейшими системами являются рекурсивные системы, в которых матрица коэффициентов при эндогенных переменных имеет треугольный вид. Такие коэффициенты идентифицируемы, и для их оценивания используется обычный метод наименьших квадратов.

4. Для оценивания коэффициентов систем одновременных уравнений в общем случае используются специальные методы: двух- и трехшаговые методы наименьших квадратов, методы неподвижной точки и др. Наиболее употребительным является двухшаговый метод наименьших квадратов, который дает состоятельные оценки, достаточно хорошие и для конечных выборок. Он применяется к каждому уравнению в отдельности и состоит в вычислении регрессии эндогенных объясняющих переменных, входящих в n-е уравнение, на все предопределенные переменные системы, а затем в использовании для оценивания искомых коэффициентов n-го уравнения вместо данных значений объясняющих переменных их оценок, полученных на первом шаге.

Глава 15. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ СТАТИСТИЧЕСКОГО ИССЛЕДОВАНИЯ ЗАВИСИМОСТЕЙ

Методы статистического исследования зависимостей, в особенности регрессионный анализ, анализ временных рядов, дисперсионный анализ, анализ таблиц сопряженности, планирование эксперимента, наиболее употребительны среди методов обработки данных в различных областях науки и техники. Соответственно к настоящему времени существует и продолжает разрабатываться обширное программное обеспечение, связанное с исследованием зависимостей. Ниже кратко рассмотрены программные средства — пакеты и библиотеки программ, доступные пользователям в СССР, а также наиболее интересные, на наш взгляд, для обеспечения статистического исследования зависимостей зарубежные пакеты и библиотеки. Основные сведения о пакетах и библиотеках программ приведены в табл. 15.1.

В табл. 15.2 приведены данные о разделах статистического исследования зависимостей, охватываемых пакетами и библиотеками программ. Номера вертикальных граф табл. 15.2 соответствуют следующим разделам и методам статистического исследования зависимостей:

1. Корреляционный анализ:

   а) вычисление и анализ значимости множеств частных и множественных коэффициентов корреляции;

   б) анализ коэффициентов корреляции Спирмэна, Кендалла и др.

2. Регрессионный анализ:

   а) линейная множественная регрессия;

   б) отбор переменных в линейной регрессии методом полного перебора или «ветвей и границ»;

   в) пошаговые процедуры отбора переменных;

   г) регрессия на главные компоненты;

   д) гребневая регрессия и другие виды регрессии в условиях мультиколлинеарности;

   е) робастная регрессия;

   ж) полиномиальная регрессия с подбором степени полинома;

   з) регрессия нелинейная относительно оцениваемых параметров;

   и) оценивание параметров стандартных нелинейных регрессионных моделей (логистическая, экспоненциальные и другие кривые);

   к) непараметрическая регрессия.

3. Дисперсионный анализ (ДА).

4. Статистическое исследование зависимостей в случае неколичественных переменных и переменных смешанной природы:

   а) стандартный анализ таблиц сопряженности;

   б) логлинейный анализ таблиц сопряженности;

   в) выравнивание шкал (оцифровка, переход к бинарным переменным).

5. Статистический анализ систем структурных эконометрических уравнений.

6. Анализ временных рядов.

Приведем теперь сведения о программном обеспечении некоторых разделов статистического исследования зависимостей, не включенные в табл. 15.1, 15.2.

Анализ временных рядов. Дополнительная информация приведена в табл. 15.3.

Логлинейный анализ таблиц сопряженности. Соответствующее программное обеспечение описано в [91, 23].

Анализ нечисловой информации. Программное обеспечение обработки нечисловой информации методами оцифровки представлено в [121], корреспонденс-анализа — в [221].

Отбор переменных в регрессионном анализе. Программа, реализующая метод «ветвей и границ» и некоторые другие ме-

Таблица 15.1

Пакет или библиотека программ	Организация-разработчик	ЭВМ	Операционная система	Использованный язык программирования	Литературные источники	Примечание
1	2	3	4	5	6	7
ПНП—БИМ (пакет научных подпрограмм, библиотека института математики АН БССР)	Институт математики АН БССР, г. Минск	ЕС ЭВМ	ОС, ДОС	Фортран	[111]	Распространяется в виде загрузочных модулей
ППСА (пакет программ по прикладному статистическому анализу)	Центральный экономико-математический институт АН СССР, г. Москва	ЕС ЭВМ (модели 1022 и выше)	ОС	ПЛ/I	[99]	Пакет программ статистической обработки многомерных данных многоцелевого назначения. Управляется с помощью языка пользователя; распространяется в виде загрузочных модулей
Система статистической обработки	Тартуский ордена Трудового Красного Знамени государственный университет г. Тарту	Минск 32 ЕС ЭВМ	ОС	Фортран	[127, 129]	Пакет программ статистической обработки данных многоцелевого назначения. Управляется с помощью языка пользователя
Пакет программ по математической статистике	Институт кибернетики АН Эст. ССР, г. Таллин	ЕС ЭВМ	ОС	Фортран	[72, 73, 74]	То же

Продолжение табл. 15.1

Пакет или библиотека программ	Организация-разработчик	ЭВМ	Операционная система	Использованный язык программирования	Литературные источники	Примечание
1	2	3	4	5	6	7
ОТЭКС (пакет прикладных программ для обработки таблиц экспериментальных данных)	Институт математики СО АН СССР, г. Новосибирск	ЕС ЭВМ	ОС	Фортран	[98]	Предназначен в основном для решения задач классификации; управляется в диалоговом режиме; для исследования зависимостей имеются нетрадиционные процедуры: — непараметрического прогноза значений независимой переменной (алгоритм ZET); — определения логических закономерностей для предсказания значений количественных (с точностью до интервала) и порядковых признаков; — отбора переменных в регрессионном анализе по величине коэффициента детерминации с помощью метода случайного поиска с адаптацией [76]

СОД—ГС (пакет прикладных программ статистической обработки данных)	Институт кибернетики АН УССР, г. Киев	ЕС ЭВМ	ОС	Фортран	[100]	Управляется с помощью языка пользователя, обрабатывающая часть построена на основе подпрограмм ПНП-БИМ
СОРРА-2 (система оперативной разработки распознающих алгоритмов)	Институт математики и кибернетики АН Лит. ССР, г. Вильнюс	БЭСМ-6	ДИСПАК	Фортран	[104]	Пакет предназначен в основном для решения задач классификации при наличии обучающих выборок; статистическое исследование зависимостей представлено программами: — линейной регрессии с возможной проверкой качества методом скользящего экзамена; — непараметрической регрессии; — пошаговой процедуры отбора переменных. Управляется с помощью языка пользователя, имеется возможность работы в полудиалоговом режиме
DIAS (диалоговая автоматизированная система)	Вычислительный центр СО АН СССР, г. Новосибирск	БЭСМ-6	DIMON	Фортран	[141]	Управляется с помощью языка пользователя, обрабатывающая часть построена на основе ПНП IBM [118]

Продолжение табл. 15.1

Пакет или библиотека программ	Организация-разработчик	ЭВМ	Операционная система	Использованный язык программирования	Литературные источники	Примечание
1	2	3	4	5	6	7
САОН (статистический анализ и обработка наблюдений)	Всесоюзный институт по проектированию организаций энергетического строительства (Оргэнергострой) Минэнерго СССР	ЕС ЭВМ	ОС	ПЛ/I	[122[	Содержит широкий спектр методов статистической обработки данных (дескриптивная статистика, непараметрическая статистика, факторный анализ, статистическое исследование зависимостей и т. д.). Управляется с помощью языка пользователя
NAG (библиотеки подпрограмм)	NAG Ltd., Великобритания, NAG (Numerical algorithms group) объединение сотрудников университетов Англии, занимающихся разработкой программного обеспечения в области численного анализа и статистики	Практически все типы ЭВМ, имеющие трансляторы с Фортрана, Алгол 60, Алгол 68		Фортран, Алгол 60, Алгол 68	[153, 158]	Существует значительное число модификаций этих библиотек, в зависимости от конкретных версий Фортрана, Алгол 60 и Алгол 68. Наиболее полные версии библиотек MARK-8, MARK-9 } Фортран содержат около 500 подпрограмм. Библиотеки распространяются в виде текстов подпрограмм на

						соответствующем языке высокого уровня (печатные тексты, микрофиши, магнитная лента), а при желании пользователя — и в виде загрузочных модулей
GENSTAT[1] (General Statistical Program)	То же	IBM 360/370 CDS DEC SYSTEM 10/20	OS, OS/VS NOS, SCOPE 20 TOPS 10/20	Фортран	[153]	Многоцелевая система статистического анализа. Управляется с помощью языка пользователя. Обрабатывающая часть создана на базе библиотек NAG
GLIM[1] (General Linear Model)	То же	То же	То же	Фортран	[158]	Программа для оценки параметров обобщенной линейной модели в интерактивном режиме. Управляется с помощью языка (команд) пользователя. Обрабатывающая часть создана на базе библиотек NAG
P—STAT[1]	Вычислительный центр Принстонского университета, США. Правом распространения пакета владеет P-STAT Inc.	IBM 360/370	OS, OS/VS, VS1	Фортран	[241]	Многоцелевая система статистического анализа. Отличительной чертой является чрезвычайно широкая возможность управления и манипуляция данными. Возможна работа в интерактивном режиме

Продолжение табл. 15.1

Пакет или библиотека программ	Организация-разработчик	ЭВМ	Операционная система	Использованный язык программирования	Литературные источники	Примечание
1	2	3	4	5	6	7
IMSL (International Mathematical Statistical Library)	IMSL Inc., США	Все типы ЭВМ, имеющие транслятор с Фортран		Фортран	[218]	Имеются многочисленные модификации библиотек IMSL, настроенные на различные «диалекты» Фортрана. Является одной из самых обширных библиотек в области численного анализа, линейной алгебры и статистического анализа. Подпрограммы, входящие в состав IMSL, обладают высокой надежностью
SPSS[1] (Statistical Package for the Social Sciences)	SPSS Inc., США	IBM 360/370 DEC SYSTEM 10/20 DEC PDP11	OS, OS/VS, VS1 TOPS 10/20 PSTS, UNIX RT-11, IAS RSX-11	Фортран	[250]	Пакет программ статистической обработки многоцелевого назначения. Управляется с помощью языка пользователя. В настоящее время разработана версия этого пакета SCSS для работы в интерактивном режиме
SAS	*SAS* Institute Inc., США	IBM 360/370	OS, OS/VS, VS1 и т. д.	ПЛ/I Ассемблер	[160]	Система программирования для статистической обработки и управления

						данными. Обладает наиболее широкими возможностями управления данными среди известных пакетов программ статистической обработки. Язык управления, помимо интегрированных команд на обработку данных и управления ими, позволяет легко включать матричные операции, организовывать циклы из процедур обработки и т. д. Возможна работа в интерактивном режиме. В отличие от большинства зарубежных пакетов предназначен только для ЭВМ типа IBM 360/370
BMDP (Biomedical Computer Programs)	**BMDP Software, США**	**IBM 360/370**	OS, OS/VS, VS1 и т. п.	Фортран	[169]	Многоцелевой пакет программ статистической обработки данных. Управляется с помощью языка пользователя. Последние версии допускают интерактивный режим. Разработаны версии пакета для весьма широкого круга ЭВМ. Версия этого пакета от 1975 г. адаптирована в СССР для ЕС ЭВМ [84]

[1] В графе 3 перечислены лишь некоторые типы ЭВМ, для которых имеются версии пакета. Более подробную информацию можно получить из соответствующего литературного источника. Версии пакета для различных ЭВМ могут отличаться по некоторым характеристикам (максимальное допустимое число переменных, разрядность операций, возможности управления данными и т. д.).

Пакет или библиотека	1		2					
	а	б	а	б	в	г	д	е
ПНП—БИМ[1]	да	да	да	нет	да	нет	нет	да
ППСА	да	да	да	нет	да	да	да	да
ССОД Тартуского ордена Трудового Красного Знамени государственного университета	да	да	да	нет	да	нет	нет	нет
ППМС Института кибернетики АН Эст. ССР	да	да	да	нет	да	нет	нет	нет
ОТЭКС	нет	нет	нет	нет[2]	нет	нет	нет	нет
СОД—ГС	да	нет	да	нет	нет	нет	нет	нет
СОРРА-2	нет	нет	да	нет	да	нет	нет	нет
DIAS	да	нет	да	нет	да	нет	нет	нет
САОН	да	нет	да	нет	да	нет	нет	да
NAG (Библиотеки подпрограмм MARK8 MARK9)	да	да	да	нет	да	да	да	да
GENSTAT	да	да	да	нет	да	да	да	да
GLIM	да	да	да	нет	да	нет	нет	да
P—STAT	да	да	да	нет	да	нет	нет	да
IMSL	да	да	да	да	да	нет	нет	нет
SPSS	да	да	да	нет	да	нет	нет	нет
SAS	да	да	да	нет	да	нет	нет	нет
BMDP[4]	да	да	да	да[5]	да	нет	нет	да

[1] Список реализованных методов взят из [111].
[2] См. графу «Примечание» в табл. 15.1.
[3] Для проведения нелинейной регрессии можно использовать процедуры поис
[4] В версиях 1982, 1983 гг. имеются программы логлинейного анализа и анализа
[5] В версии 1975 г. отсутствует.

Таблица 15.2

				3	4			5	6
ж	з	и	к		а	б	в		
да	да	да	нет	да	нет	нет	нет	да	да
нет	да	да	нет	нет	да	нет	да	нет	нет
да	нет	нет	нет	да	да	нет	да	нет	нет
да	нет	нет	нет	да	нет	нет	нет	нет	да
нет	нет	нет	да[2]	нет	нет	нет	да	нет	нет
да	нет	да	нет	нет	нет	нет	нет	нет	да
нет	нет	нет	да	нет	нет	нет	нет	нет	нет
нет	нет	нет	нет	да	да	нет	нет	нет	нет
да	да	нет	да	нет	нет	нет	нет	нет	да
да	нет[3]	да	нет	да	да			да	да
да	нет	да	нет	да	да	да	нет	нет	да
да	нет	да	нет	да	да	да	нет	нет	нет
да	да	да	нет	да	да	нет	нет	нет	нет
да	нет[3]	да	нет	да	да	да	нет	да	да
да	нет	нет	нет	да	да	нет	нет	нет	нет
нет	да	да	нет	да	да	нет	да	да	да
да	да	да	нет	да	да	нет	нет	нет	нет

ка минимума функции многих переменных из этой библиотеки.
временных рядов.

Таблица 15.3

Пакет или библиотека программ	Организация-разработчик	ЭВМ	Операционная система	Язык программирования	Литературные источники
Пакет промышленных программ для анализа и прогноза временных рядов	РВЦ ЦСУ РСФСР, г. Москва	ЕС ЭВМ	ОС		[101]
ПАРИСС	Киевский ордена Ленина государственный университет им. Т. Г. Шевченко, г. Киев	ЕС ЭВМ (модели 1022 и выше)	ОС	Фортран	[120]
МАВР	То же	ЕС ЭВМ (модели 1045 и выше)	ОС	»	[120]
Пакет ДЕЛЬТА-СТАТ (Δ-стат) для обработки данных в интерактивном режиме	Институт кибернетики АН УССР, г. Киев	ЕС ЭВМ	ОС	Фортран Ассемблер	[36]
TSA	NAG, Великобритания	Широкий круг ЭВМ	—	Фортран	[254]

тоды отбора, описана в [87]. Отбор в условиях многомерной функции отклика рассматривается в [151].

Математическое обеспечение прикладного статистического анализа для мини- и микроЭВМ рассматривается в [83, 157].

Дополнительную информацию по вопросу программного обеспечения статистического анализа можно получить в [24, 69, 188].

ПРИЛОЖЕНИЯ. МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ ТАБЛИЦЫ

Таблица П.1

Значения функции плотности $\varphi(x)=\varphi(x;\ 0;\ 1)$ стандартного нормального закона распределения

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$	x	$\varphi(x)$
0,00	0,3989	0,50	0,3521	1,00	0,2420	1,50	0,1295	2,00	0,0540	2,50	0,0175
0,05	0,3984	0,55	0,3429	1,05	0,2299	1,55	0,1200	2,05	0,0488	2,55	0,0154
0,10	0,3970	0,60	0,3332	1,10	0,2179	1,60	0,1109	2,10	0,0440	2,60	0,0136
0,15	0,3945	0,65	0,3230	1,15	0,2059	1,65	0,1023	2,15	0,0396	2,65	0,0119
0,20	0,3910	0,70	0,3123	1,20	0,1942	1,70	0,0940	2,20	0,0355	2,70	0,0104
0,25	0,3867	0,75	0,3011	1,25	0,1826	1,75	0,0863	2,25	0,0317	2,75	0,0091
0,30	0,3814	0,80	0,2897	1,30	0,1714	1,80	0,0790	2,30	0,0283	2,80	0,0079
0,35	0,3752	0,85	0,2780	1,35	0,1604	1,85	0,0721	2,35	0,0252	2,85	0,0069
0,40	0,3683	0,90	0,2661	1,40	0,1497	1,90	0,0656	2,40	0,0224	2,90	0,0060
0,45	0,3605	0,95	0,2541	1,45	0,1394	1,95	0,0596	2,45	0,0198	2,95	0,0051
										3,00	0,0044

Пример. Требуется определить значение функции плотности $\varphi(x;\ a;\ \sigma)$ нормального закона в точке $x = 3{,}36$ при среднем $a = 1$ и среднеквадратическом отклонении $\sigma = 2$.

1. Подсчитываем $x_0 = (x - a)/\sigma = (3{,}36 - 1)/2 = 1{,}18$.

2. Находим с помощью таблицы значение функции $\varphi(x;\ 0;\ 1)$ в точке $x_0 = 1{,}18$, прибегая в случае необходимости к линейной интерполяции, а именно:

$$\varphi(x_0;\ 0;\ 1)=\varphi(x_1;\ 0;\ 1)-\frac{x_0-x_1}{x_2-x_1}\left[\varphi(x_1;\ 0;\ 1)-\varphi(x_2;\ 0;\ 1)\right],$$

где x_1 и x_2 — два соседних табличных значения аргумента ($x_1 < x_2$), между которыми находится интересующее нас значение x_0. В нашем примере

$$\varphi(1{,}18;\ 0;\ 1)=\varphi(1{,}15;\ 0;\ 1)-\frac{1{,}18-1{,}15}{1{,}20-1{,}15}\left[\varphi(1{,}15)-\varphi(1{,}20)\right]=$$

$$=0{,}2059-\frac{3}{5}(0{,}2059-0{,}1942)=0{,}199.$$

3. Подсчитываем $\varphi(x;\ a;\ \sigma)$ с помощью формулы

$$\varphi(x;\ a;\ \sigma)=\frac{1}{\sigma}\varphi\left(\frac{x-a}{\sigma};\ 0;\ 1\right)=\frac{1}{2}\varphi(1{,}18;\ 0;\ 1)=0{,}0995.$$

Замечание. В силу четности функции $\varphi(x)$ для нахождения ее значений при отрицательных величинах аргумента достаточно найти из таблицы ее значение при том же самом, но положительном аргументе [т. е. $\varphi(-x_0) \equiv \varphi(x_0)$ при любом x_0].

Таблица П.2

Значения функции $\Phi(x)=\Phi(x;\ 0;\ 1)$ стандартного нормального распределения

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-t^2/2}\,dt$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,500000	1,00	0,841345	2,00	0,977250
0,05	0,519939	1,05	0,853141	2,05	0,979818
0,10	0,539828	1,10	0,864334	2,10	0,982136
0,15	0,559618	1,15	0,874928	2,15	0,984222
0,20	0,579260	1,20	0,884930	2,20	0,986097
0,25	0,589706	1,25	0,894350	2,25	0,987776
0,30	0,617911	1,30	0,903200	2,30	0,989276
0,35	0,636831	1,35	0,911492	2,35	0,990613
0,40	0,655422	1,40	0,919243	2,40	0,991802
0,45	0,673645	1,45	0,926471	2,45	0,992857
0,50	0,691463	1,50	0,933193	2,50	0,993790
0,55	0,708840	1,55	0,939429	2,55	0,994614
0,60	0,725747	1,60	0,945201	2,60	0,995339
0,65	0,742154	1,65	0,950528	2,65	0,995975
0,70	0,758036	1,70	0,955434	2,70	0,996533
0,75	0,773373	1,75	0,959941	2,75	0,997020
0,80	0,788145	1,80	0,964070	2,80	0,997445
0,85	0,802338	1,85	0,967843	2,85	0,997814
0,90	0,815940	1,90	0,971283	2,90	0,998134
0,95	0,828944	1,95	0,974412	2,95	0,998411
				3,00	0,998650

Пример. Требуется определить значение функции распределения $\Phi(x;\ a;\ \sigma)$ нормального закона в точке $x = 3{,}36$ при среднем $a = 1$ и среднеквадратическом отклонении $\sigma = 2$.

1. Подсчитываем $x_0 = (x - a)/\sigma = (3{,}36 - 1)/2 = 1{,}18$.

2. Находим с помощью таблицы значение функции $\Phi(x)$ в точке $x_0 = 1{,}18$, прибегая в случае необходимости к линейной интерполяции, а именно:

$$\Phi(x_0)=\Phi(x_1)+\frac{x_0-x_1}{x_2-x_1}\left[\Phi(x_2)-\Phi(x_1)\right],$$

где x_1 и x_2 — два соседних табличных значения аргумента ($x_1 < x_2$), между которыми находится интересующее нас значение x_0. В нашем примере

$$\Phi(1{,}18)=\Phi(1{,}15)+\frac{1{,}18-1{,}15}{1{,}20-1{,}15}\left[\Phi(1{,}20)-\Phi(1{,}15)\right]=$$

$$=0{,}8749+\frac{3}{5}(0{,}8849-0{,}8749)=0{,}8809$$

З а м е ч а н и е. При отыскании значений Ф (x) для $x < 0$ следует пользоваться соотношением Ф (x) = 1 — Ф ($-x$). Например, если надо найти значение Ф (x) в точке $x = -0,25$, то имеем: Ф ($-x$) = Ф (0,25) = 0,5897, так что Ф (— 0,25) = 1 — Ф (0,25) = 1 — 0,5897 = 0,4103.

Таблица П.3

Значения q-квантилей u_q стандартного нормального распределения

q	u_q	q	u_q	q	u_q	q	u_q
0,50	0,000000	0,70	0,524401	0,90	1,281552	0,983	120072
51	025069	71	553385	91	340755	984	144411
52	050154	72	582842	92	405072	0,985	2,170090
53	075270	73	612813	93	475791	986	197286
54	100434	74	643345	94	554774	987	226212
0,55	0,125661	0,75	0,674490	0,95	1,644854	988	257129
56	150969	76	706303	96	750686	989	290368
57	176374	77	738847	97	880794	0,990	2,326348
58	201893	78	772193	971	895698	991	365618
59	227545	79	806421	972	911036	992	408916
0,60	0,253347	0,80	0,841621	973	926837	993	457263
61	279319	81	877896	974	943134	994	512144
62	305481	82	915365	0,975	1,959964	0,995	2,575829
63	331853	83	954165	976	977368	996	652070
64	358459	84	994458	977	995393	997	747781
0,65	0,385320	0,85	1,036433	978	2,014091	998	878162
66	412463	86	080319	979	033520	999	3,090232
67	439913	87	126391	0,980	2,053749		
68	467699	88	174987	981	074855		
69	495850	89	226528	982	096927		

П р и м е р. Найти 0,9-квантиль $u_{0,9}$. Величину $u_{0,9}$ находим из таблицы в графе, расположенной справа от соответствующего значения $q = 0,9$, т. е. $u_{0,9} = 1,28155$.

З а м е ч а н и е 1. Если заданная величина q попадает между двумя соседними табличными значениями q_1 и q_2 ($q_1 < q_2$; это может случиться при графической проверке нормальности распределения), то следует воспользоваться линейной интерполяцией, а именно формулой

$$u_q = u_{q_1} + \frac{q - q_1}{q_2 - q_1} (u_{q_2} - u_{q_1}).$$

З а м е ч а н и е 2. При нахождении q-квантилей для значений $q < 0,5$ следует воспользоваться соотношением $u_q = -u_{1-q}$. Например, $u_{0,4} = -u_{1-0,4} = -u_{0,6} = -0,25335$.

З а м е ч а н и е 3. При отыскании $100Q\%$-ных точек w_Q следует воспользоваться соотношением $w_Q = u_{1-Q}$. Например, $w_{0,05} = u_{0,95} = 1,64485$.

Таблица П.4

Значения $100Q\%$-ных точек $\chi^2_Q(\nu)$ χ^2-распределения с ν степенями свободы

ν	Q									
	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	$392704\cdot10^{-10}$	$157088\cdot10^{-9}$	$982069\cdot10^{-9}$	$393214\cdot10^{-8}$	0,0157908	2,70554	3,84146	5,02389	6,63490	7,87944
2	0,0100251	0,0201007	0,0506356	0,102587	0,210720	4,60517	5,99147	7,37776	9,21034	10,5966
3	0,0717212	0,114832	0,215795	0,351846	0,584375	6,25139	7,81473	9,34840	11,3449	12,8381
4	0,206990	0,297110	0,484419	0,710721	1,063623	7,77944	9,48773	11,1433	13,2767	14,8602
5	0,411740	0,554300	0,831211	1,145476	1,61031	9,23635	11,0705	12,8325	15,0863	16,7496
6	0,675727	0,872085	1,237347	1,63539	2,20413	10,6446	12,5916	14,4494	16,8119	18,5476
7	0,989265	1,239043	1,68987	2,16735	2,83311	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,344419	1,646482	2,17973	2,73264	3,48954	13,3616	15,5073	17,5346	20,0902	21,9550
9	1,734926	2,087912	2,70039	3,32511	4,16816	14,6837	16,9190	19,0228	21,6660	23,5893
10	2,15585	2,55821	3,24697	3,94030	4,86518	15,9871	18,3070	20,4831	23,2093	25,1882
11	2,60321	3,05347	3,81575	4,57481	5,57779	17,2750	19,6751	21,9200	24,7250	26,7569
12	3,07382	3,57056	4,40379	5,22603	6,30380	18,5494	21,0261	23,3367	26,2170	28,2995
13	3,56503	4,10691	5,00874	5,89186	7,04150	19,8119	22,3621	24,7356	27,6883	29,8194
14	4,07468	4,66043	5,62872	6,57063	7,78953	21,0642	23,6848	26,1190	29,1413	31,3193
15	4,60094	5,22935	6,26214	7,26094	8,54675	22,3072	24,9958	27,4884	30,5779	32,8013
16	5,14224	5,81221	6,90766	7,96164	9,31223	23,5418	26,2962	28,8454	31,9999	34,2672
17	5,69724	6,40776	7,56418	8,67176	10,0852	24,7690	27,5871	30,1910	33,4087	35,7185
18	6,26481	7,01491	8,23075	9,39046	10,8649	25,9894	28,8693	31,5264	34,8053	37,1564
19	6,84398	7,63273	8,90655	10,1170	11,6509	27,2036	30,1435	32,8523	36,1908	38,5822

20	7,43386	8,26040	9,59083	10,8508	12,4426	28,4120	31,4104	34,1696	37,5662	39,9968
21	8,03366	8,89720	10,28293	11,5913	13,2396	29,6151	32,6705	35,4789	38,9321	41,4010
22	8,64272	9,54249	10,9823	12,3380	14,0415	30,8133	33,9244	36,7807	40,2894	42,7956
23	9,26042	10,19567	11,6885	13,0905	14,8479	32,0069	35,1725	38,0757	41,6384	44,1813
24	9,88623	10,8564	12,4011	13,8484	15,6587	33,1963	36,4151	39,3641	42,9798	45,5585
25	10,5197	11,5240	13,1197	14,6114	16,4734	34,3816	37,6525	40,6465	44,3141	46,9278
26	11,1603	12,1981	13,8439	15,3791	17,2919	35,5631	38,8852	41,9232	45,6417	48,2899
27	11,8076	12,8786	14,5733	16,1513	18,1138	36,7412	40,1133	43,1944	46,9630	49,6449
28	12,4613	13,5648	15,3079	16,9279	18,9392	37,9159	41,3372	44,4607	48,2782	50,9933
29	13,1211	14,2565	16,0471	17,7083	19,7677	39,0875	42,5569	45,7222	49,5879	52,3356
30	13,7867	14,9535	16,7908	18,4926	20,5992	40,2560	43,7729	46,9792	50,8922	53,6720
40	20,7065	22,1643	24,4331	26,5093	29,0505	51,8050	55,7585	59,3417	63,6907	66,7659
50	27,9907	29,7067	32,3574	34,7642	37,6886	63,1671	67,5048	71,4202	76,1539	79,4900
60	35,5346	37,4848	40,4817	43,1879	46,4589	74,3970	79,0819	83,2976	88,3794	91,9517
70	43,2752	45,4418	48,7576	51,7393	55,3290	85,5271	90,5312	95,0231	100,425	104,215
80	51,1720	53,5400	57,1532	60,3915	64,2778	96,5782	101,879	106,629	112,329	116,321
90	59,1963	61,7541	65,6466	69,1260	73,2912	107,565	113,145	118,136	124,116	128,299
100	67,3276	70,0648	74,2219	77,9295	82,3581	118,498	124,342	129,561	135,807	140,169

Таблица П.5

Значения 100 Q%-ных точек v_Q^2 (ν_1, ν_2) F-распределения с числом степеней свободы числителя ν_1 и знаменателя ν_2

ν_2	ν_1																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
	$Q=0,1$																		
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06	63,33
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53

25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00

$Q = 0,05$

1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13

Продолжение табл. 5

ν_2	ν_1 = 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2 27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

$Q = 0,01$

ν_2	ν_1 = 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46

5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Примечание. При вычислении $100Q$%-ных точек для $Q \geqslant 0{,}9$ следует воспользоваться тождеством $v_Q^2(\nu_1, \nu_2) = \left(v_{1-Q}^2(\nu_1, \nu_2)\right)^{-1}$.

Таблица П6

Значения $100Q\%$-ных точек $t_Q(\nu)$ распределения Стьюдента (t-распределения) с ν степенями свободы

ν	$Q=0,4$ $2Q=0,8$	0,25 0,5	0,1 0,2	0,05 0,1	0,025 0,05	0,01 0,02	0,005 0,01	0,0025 0,005
1	0,325	1,000	3,078	6,314	12,706	31,821	63,657	127,32
2	289	0,816	1,886	2,920	4,303	6,965	9,925	14,089
3	277	765	1,638	2,353	3,182	4,541	5,841	7,453
4	271	741	1,533	2,132	2,776	3,747	4,604	5,598
5	0,267	0,727	1,476	2,015	2,571	3,365	4,032	4.773
6	265	718	1,440	1,943	2,447	3,143	3,707	4,317
7	263	711	1,415	1,895	2,365	2,998	3,499	4,029
8	262	706	1,397	1,860	2,306	2,896	3,355	3,833
9	261	703	1,383	1,833	2,262	2,821	3,250	3,690
10	0,260	0,700	1,372	1,812	2,228	2,764	3,169	3,581
11	260	697	1,363	1,796	2,201	2,718	3,106	3,497
12	259	695	1,356	1,782	2,179	2,681	3,055	3,428
13	259	694	1,350	1,771	2,160	2,650	3,012	3,372
14	258	692	1,345	1,761	2,145	2,624	2,977	3,326
15	0,258	0,691	1,341	1,753	2,131	2,602	2,947	3,286
16	258	690	1,337	1,746	2,120	2,583	2,921	3,252
17	257	689	1,333	1,740	2,110	2,567	2,898	3,222
18	257	688	1,330	1,734	2,101	2,552	2,878	3,197
19	257	688	1,328	1,729	2,093	2,539	2,861	3,174
20	0,257	0,687	1,325	1,725	2,086	2,528	2,845	3,153
21	257	686	1,323	1,721	2,080	2,518	2,831	3,135
22	256	686	1,321	1,717	2,074	2,508	2,819	3,119
23	256	685	1,319	1,714	2,069	2,500	2,807	3,104
24	256	685	1,318	1,711	2,064	2,492	2,797	3,091
25	0,256	0,684	1,316	1,708	2,060	2,485	2,787	3,078
26	256	684	1,315	1,706	2,056	2,479	2,779	3,067
27	256	684	1,314	1,703	2,052	2,473	2,771	3,057
28	256	683	1,313	1,701	2,048	2,467	2,763	3,047
29	256	683	1,311	1,699	2,045	2,462	2,756	3,038
30	0,256	0,683	1,310	1,697	2,042	2,457	2,750	3,030
40	255	681	1,303	1,684	2,021	2,423	2,704	2.971
60	254	679	1,296	1,671	2,000	2,390	2,660	2,915
120	254	677	1,289	1,658	1,980	2,358	2,617	2,860
∞	253	674	1,282	1,645	1,960	2,326	2,576	2,807

Таблица П.7

Преобразование Фишера (z-преобразование) выборочного коэффициента корреляции $\hat{r}$
($\hat{r} = \text{th}\, z$, $z = \text{arc th}\, \hat{r}$)

$\hat{r}$	,000	,002	,004	,006	,008	$\hat{r}$	,000	,002	,004	,006	,008
0,00	0000	0020	0040	0060	0080	0,50	5493	5520	5547	5573	5600
1	0100	0120	0140	0160	0180	1	5627	5654	5682	5709	5736
2	0200	0220	0240	0260	0280	2	5763	5791	5818	5846	5874
3	0300	0320	0340	0360	0380	3	5901	5929	5957	5985	6013
4	0400	0420	0440	0460	0480	4	6042	6070	6098	6127	6155
0,05	0500	0520	0541	0561	0581	0,55	6184	6213	6241	6270	6299
6	0601	0621	0641	0661	0681	6	6328	6358	6387	6416	6446
7	0701	0721	0741	0761	0782	7	6475	6505	6535	6565	6595
8	0802	0822	0842	0862	0882	8	6625	6655	6685	6716	6746
9	0902	0923	0943	0963	0983	9	6777	6807	6838	6869	6900
0,10	1003	1024	1044	1064	1084	0,60	6931	6963	6994	7026	7057
1	1104	1125	1145	1165	1186	1	7089	7121	7153	7185	7218
2	1206	1226	1246	1267	1287	2	7250	7283	7315	7348	7381
3	1307	1328	1348	1368	1389	3	7414	7447	7481	7514	7548
4	1409	1430	1450	1471	1491	4	7582	7616	7650	7684	7718
0,15	1511	1532	1552	1573	1593	0,65	7753	7788	7823	7858	7893
6	1614	1634	1655	1676	1696	6	7928	7964	7999	8035	8071
7	1717	1737	1758	1779	1799	7	8107	8144	8180	8217	8254
8	1820	1841	1861	1882	1903	8	8291	8328	8366	8404	8441
9	1923	1944	1965	1986	2007	9	8480	8518	8556	8595	8634
0,20	2027	2048	2069	2090	2111	0,70	8673	8712	8752	8792	8832
1	2132	2153	2174	2195	2216	1	8872	8912	8953	8994	9035
2	2237	2258	2279	2300	2321	2	9076	9118	9160	9202	9245
3	2342	2363	2384	2405	2427	3	9287	9330	9373	9417	9461
4	2448	2469	2490	2512	2533	4	9505	9549	9594	9639	9684
0,25	2554	2575	2597	2618	2640	0,75	0,973	0,978	0,982	0,987	0,991
6	2661	2683	2704	2726	2747	6	0,996	1,001	1,006	1,011	1,015
7	2769	2790	2812	2833	2855	7	1,020	1,025	1,030	1,035	1,040
8	2877	2899	2920	2942	2964	8	1,045	1,050	1,056	1,061	1,066
9	2986	3008	3029	3051	3073	9	1,071	1,077	1,082	1,088	1,093
0,30	3095	3117	3139	3161	3183	0,80	1,099	1,104	1,110	1,116	1,121
1	3205	3228	3250	3272	3294	1	1,127	1,133	1,139	1,145	1,151
2	3316	3339	3361	3383	3406	2	1,157	1,163	1,169	1,175	1,182
3	3428	3451	3473	3496	3518	3	1,188	1,195	1,201	1,208	1,214
4	3541	3564	3586	3609	3632	4	1,221	1,228	1,235	1,242	1,249
0,35	3654	3677	3700	3723	3746	0,85	1,256	1,263	1,271	1,278	1,286
6	3769	3792	3815	3838	3861	6	1,293	1,301	1,309	1,317	1,325
7	3884	3907	3931	3954	3977	7	1,333	1,341	1,350	1,358	1,367
8	4001	4024	4047	4071	4094	8	1,376	1,385	1,394	1,403	1,412
9	4118	4142	4165	4189	4213	9	1,422	1,432	1,442	1,452	1,462

$\hat{r}$	,000	,002	,004	,006	,008	$\hat{r}$	,000	,002	,004	,006	,008
0,40	4236	4260	4284	4308	4332	0,90	1,472	1,483	1,494	1,505	1,516
1	4356	4380	4404	4428	4453	1	1,528	1,539	1,551	1,564	1,576
2	4477	4501	4526	4550	4574	2	1,589	1,602	1,616	1,630	1,644
3	4599	4624	4648	4673	4698	3	1,658	1,673	1,689	1,705	1,721
4	4722	4747	4772	4797	4822	4	1,738	1,756	1,774	1,792	1,812
0,45	4847	4872	4897	4922	4948	0,95	1,832	1,853	1,874	1,897	1,921
6	4973	4999	5024	5049	5075	6	1,946	1,972	2,000	2,029	2,060
7	5101	5126	5152	5178	5204	7	2,092	2,127	2,165	2,205	2,249
8	5230	5256	5282	5308	5334	8	2,298	2,351	2,410	2,477	2,555
9	5361	5387	5413	5440	5466	9	2,647	2,759	2,903	3,106	3,453
$\hat{r}$	,000	,002	,004	,006	,008	$\hat{r}$	,000	,002	,004	,006	,008

Примеры.

1. Дано $\hat{r} = 0{,}206$. Определить $z = \operatorname{arc\,th} 0{,}206$.

Находим (в левом столбце таблицы) строку, соответствующую $\hat{r} = 0{,}20$. Чтобы получить заданное значение $\hat{r}$, к 0,20 надо прибавить 0,006, а потому искомое число находится (в этой строке) в столбце, расположенном под 0,006. Итак, $z = \operatorname{arc\,th} 0{,}206 = 0{,}2090$.

2. Дано $\hat{r} = -0{,}515$. Определить $z = \operatorname{arcth} (-0{,}515)$.

Находим (в левом столбце таблицы) строку, соответствующую $\hat{r} = 0{,}51$. Чтобы получить значение $\hat{r} = 0{,}515$, к 0,51 надо прибавить 0,005, а потому arc th 0,515 находится как среднее арифметическое двух чисел данной строки, расположенных в столбцах, соответствующих верхним индексам 0,006 и 0,004, т. е.

$$\operatorname{arc\,th} 0{,}515 = \frac{0{,}5709 + 0{,}5682}{2} = 0{,}56955.$$

Соответственно $z = \operatorname{arc\,th} (-0{,}515) = -\operatorname{arcth} 0{,}515 = -0{,}56955$.

3. Дано $z = 0{,}8752$. Определить $\hat{r} = \operatorname{th} z$.

Находим в таблице число, равное 0,8752, и определяем, какому значению $\hat{r}$ оно соответствует. В нашем случае $\hat{r} = 0{,}704$.

Примечание. В тех случаях, когда в таблице не найдется в точности заданного числа, берут два приближенных (ближайших к нему) значения — с недостатком и с избытком. Искомое значение $\hat{r}$ будет лежать между двумя значениями $\hat{r}_1$ и $\hat{r}_2$, соответствующими этим приближенным величинам z.

Таблица П.8

Верхняя (положительная) граница доверительного интервала для истинного значения коэффициента корреляции в случае отсутствия корреляционной связи (при доверительной вероятности $P=1-2Q$)

$n-2$	Q					
	0,05	0,025	0,01	0,005	0,0025	0,0005
1	0,9877	$0,9^2692$	$0,9^3507$	$0,9^3877$	$0,9^4692$	$0,9^5877$
2	9000	9500	9800	9^2000	9^2500	9^3000
3	805	878	9343	9587	9740	9^2114
4	729	811	882	9172	9417	9741
5	669	754	833	875	9056	9509
6	0,621	0,707	0,789	0,834	0,870	0,9249
7	582	666	750	798	836	898
8	549	632	715	765	805	872
9	521	602	685	735	776	847
10	497	576	658	708	750	823
11	0,476	0,553	0,634	0,684	0,726	0,801
12	457	532	612	661	703	780
13	441	514	592	641	683	760
14	426	497	574	623	664	742
15	412	482	558	606	647	725
16	0,400	0,468	0,543	0,590	0,631	0,708
17	389	456	529	575	616	693
18	378	444	516	561	602	679
19	369	433	503	549	589	665
20	360	423	492	537	576	652
25	0,323	0,381	0,445	0,487	0,524	0,597
30	296	349	409	449	484	554
35	275	325	381	418	452	519
40	257	304	358	393	425	490
45	243	288	338	372	403	465
50	0,231	0,273	0,322	0,354	0,384	0,443
60	211	250	295	325	352	408
70	195	232	274	302	327	380
80	183	217	257	283	307	357
90	173	205	242	267	290	338
100	164	195	230	254	276	321

Примечание. Верхний индекс (2,3 и т. д.) над цифрой 9 означает, что эта цифра занимает первые 2,3 и т. д. разряда десятичной дроби. Например, $0,9^4692 = 0,9999692$.

Пример. Если мы оцениваем корреляционную связь по $n = 20$ наблюдениям, то при доверительной вероятности $P = 0,95$ (т. е. при $2Q = 0,05$) значение коэффициента корреляции, не превосходящее по абсолютной величине 0,444, еще не говорит о статистической значимости этой корреляционной связи (т. е. о том, что истинное значение коэффициента корреляции r отлично от нуля).

Таблица П.9

Проверка статистической значимости корреляционной связи с помощью рангового коэффициента корреляции Спирмэна $\hat{\tau}(S)$

$n=4$		$n=5$		$n=6$		$n=7$		$n=8$		$n=9$		$n=10$	
S_C	Q	S_C	Q	S_C	Q	S_C	Q	S_C	Q	S_C	Q	S_C	Q
12	0,458	22	0,475	50	0,210	74	0,249	108	0,250	156	0,218	208	0,235
14	375	24	392	52	178	78	198	114	195	164	168	218	184
16	208	26	342	54	149	82	151	120	150	172	125	228	139
18	167	28	258	56	121	86	118	126	108	180	089	238	102
20	042	30	225	58	088	90	083	132	076	188	060	248	072
		32	0,175	60	0,068	94	0.055	138	0,048	196	0,038	258	0,048
		34	117	62	051	98	033	144	029	204	022	268	030
		36	067	64	029	102	017	150	014	212	011	278	017
		38	042	66	017	106	0062	156	0054	220	0041	288	0087
		40	0083	68	0083	110	0014	162	0011	228	0010	298	0036
				70	0,0014							308	0,001
20		40		70		112		168		240		330	

Таблица П.10

Проверка статистической значимости корреляционной связи с помощью рангового коэффициента корреляции Кендалла $\hat{\tau}^{(K)}$

S_K	$n=4$	$n=5$	$n=8$	$n=9$	S_K	$n=6$	$n=7$	$n=10$
0	0,625	0,592	0,548	0,540	1	0,500	0,500	0,500
2	375	408	452	460	3	360	386	431
4	167	242	360	381	5	235	281	364
6	042	117	274	306	7	136	191	300
8		042	199	238	9	068	119	242
10		0,0083	0,138	0,179	11	0,028	0,068	0,190
12			089	130	13	0083	035	146
14			054	090	15	0014	015	108
16			031	060	17		0054	078
18			016	038	19		0014	054
20			0,0071	0,022	21		0,0002	0,036
22			0028	012	23			023
24			0009	0063	25			014
26			0002	0029	27			0083
28				0012	29			0046
30				0,0004	31			0,0023
					33			0011
					35			0005

Таблица П.11[a]

Проверка статистической значимости выборочного значения коэффициента конкордации $\hat{w}(m)$

Вероятность того, что данное значение S будет достигнуто или превзойдено, для $n=3$ и m от 2 до 10

S	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$	$m=9$	$m=10$
0	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
2	0,833	0,944	0,931	0,954	0,956	0,964	0,967	0,971	0,974
6	0,500	0,528	0,653	0,691	0,740	0,768	0,794	0,814	0,830
8	0,167	0,361	0,431	0,522	0,570	0,620	0,654	0,685	0,710
14		0,194	0,273	0,367	0,430	0,486	0,531	0,569	0,601
18		0,028	0,125	0,182	0,252	0,305	0,355	0,398	0,436
24			0,069	0,124	0,184	0,237	0,285	0,328	0,368
26			0,042	0,093	0,142	0,192	0,236	0,278	0,316
32			0,0046	0,039	0,072	0,112	0,149	0,187	0,222
38				0,024	0,052	0,085	0,120	0,154	0,187
42				0,0085	0,029	0,051	0,079	0,107	0,135
50				$0{,}0^{3}77$	0,012	0,027	0,047	0,069	0,092
54					0,0081	0,021	0,038	0,057	0,078
56					0,0055	0,016	0,030	0,048	0,066
62					0,0017	0,0084	0,018	0,031	0,046
72					$0{,}0^{3}13$	0,0036	0,0099	0,019	0,030
74						0,0027	0,0080	0,016	0,026
78						0,0012	0,0048	0,010	0,018
86						$0{,}0^{3}32$	0,0024	0,0060	0,012
96						$0{,}0^{3}32$	0,0011	0,0035	0,0075
98						$0{,}0^{4}21$	$0{,}0^{3}86$	0,0029	0,0063
104							$0{,}0^{3}26$	0,0013	0,0034
114							$0{,}0^{4}61$	$0{,}0^{3}66$	0,0020
122							$0{,}0^{4}61$	$0{,}0^{3}35$	0,0013
126							$0{,}0^{4}61$	$0{,}0^{3}20$	$0{,}0^{3}83$
128							$0{,}0^{5}36$	$0{,}0^{4}97$	$0{,}0^{3}51$
134								$0{,}0^{4}54$	$0{,}0^{3}37$
146								$0{,}0^{4}11$	$0{,}0^{3}18$
150								$0{,}0^{4}11$	$0{,}0^{3}11$
152								$0{,}0^{4}11$	$0{,}0^{4}85$
158								$0{,}0^{4}11$	$0{,}0^{4}44$
162								$0{,}0^{6}60$	$0{,}0^{4}20$
168									$0{,}0^{4}11$
182									$0{,}0^{5}21$
200									$0{,}0^{7}99$

Продолжение

Вероятность того, что данное значение S будет достигнуто или превзойдено, для $n=4$, $m=3$ и $m=5$

S	$m=3$	$m=5$	S	$m=5$
1	1,000	1,000	61	0,055
3	0,958	0,975	65	0,044
5	0,910	0,944	67	0,034
9	0,727	0,857	69	0,031
11	0,608	0,771	73	0,023
13	0,524	0,709	75	0,020
17	0,446	0,652	77	0,017
19	0,342	0,561	81	0,012
21	0,300	0,521	83	0,0087
25	0,207	0,445	85	0,0067
27	0,175	0,408	89	0,0055
29	0,148	0,372	91	0,0031
33	0,075	0,298	93	0,0023
35	0,054	0,260	97	0,0018
37	0,033	0,226	99	0,0016
41	0,017	0,210	101	0,0014
43	0,0017	0,162	105	$0{,}0^{3}64$
45	0,0017	0,141	107	$0{,}0^{3}33$
49		0,123	109	$0{,}0^{3}21$
51		0,107	113	$0{,}0^{3}14$
53		0,093	117	$0{,}0^{4}48$
57		0,075	125	$0{,}0^{5}30$
59		0,067		

Продолжение

Вероятность того, что данное значение S будет достигнуто или превзойдено, для $n=4$ и $m=2$, $m=4$ и $m=6$

S	$m=2$	$m=4$	$m=6$	S	$m=6$
0	1,000	1,000	1,000	82	0,035
2	0,958	0,992	0,996	84	0,032
4	0,833	0,928	0,957	86	0,029
6	0,792	0,900	0,940	88	0,023
8	0,625	0,800	0,874	90	0,022
10	0,542	0,754	0,844	94	0,017
12	0,458	0,677	0,789	96	0,014
14	0,375	0,649	0,772	98	0,013
16	0,208	0,524	0,679	100	0,010
18	0,167	0,508	0,668	102	0,0096
20	0,042	0,432	0,609	104	0,0085
22		0,389	0,574	106	0,0073
24		0,355	0,541	108	0,0061

Продолжение

S	$m=2$	$m=4$	$m=6$	S	$m=6$
26		0,324	0,512	110	0,0057
30		0,242	0,431	114	0,0040
32		0,200	0,386	116	0,0033
34		0,190	0,375	118	0,0028
36		0,158	0,338	120	0,0023
38		0,141	0,317	122	0,0020
40		0,105	0,270	126	0,0015
42		0,094	0,256	128	0,0^{3}90
44		0,077	0,230	130	0,0^{3}87
46		0,068	0,218	132	0,0^{3}73
48		0,054	0,197	134	0,0^{3}65
50		0,052	0,194	136	0,0^{3}40
52		0,036	0,163	138	0,0^{3}36
54		0,033	0,155	140	0,0^{3}28
56		0,019	0,127	144	0,0^{3}24
58		0,014	0,114	146	0,0^{3}22
62		0,012	0,108	148	0,0^{3}12
64		0,0069	0,089	150	0,0^{4}95
66		0,0062	0,088	152	0,0^{4}62
68		0,0027	0,073	154	0,0^{4}46
70		0,0027	0,066	158	0,0^{4}24
72		0,0016	0,060	160	0,0^{4}16
74		0,0^{3}94	0,056	162	0,0^{4}12
76		0,0^{3}94	0,043	164	0,0^{5}80
78		0,0^{3}94	0,041	170	0,0^{5}24
80		0,0^{4}72	0,037	180	0,0^{6}13

Продолжение

Вероятность того, что данное значение S будет достигнуто или превзойдено, для $n=5$ и $m=3$

S	$m=3$	S	$m=3$	S	$m=3$	S	$m=3$
0	1,000	22	0,649	44	0,236	66	0,038
2	1,000	24	0,595	46	0,213	68	0,028
4	0,988	26	0,559	48	0,172	70	0,026
6	0,972	28	0,493	50	0,163	72	0,017
8	0,941	30	0,475	52	0,127	74	0,015
10	0,914	32	0,432	54	0,117	76	0,0078
12	0,845	34	0,406	56	0,096	78	0,0053
14	0,831	36	0,347	58	0,080	80	0,0040
16	0,768	38	0,326	60	0,063	82	0,0028
18	0,720	40	0,291	62	0,056	86	0,0^{3}90
20	0,682	42	0,253	64	0,045	90	0,0^{4}69

Таблица П.11[б]

Проверка статистической значимости выборочного значения коэффициента конкордации $\hat{W}(m)$

Критические значения S при уровне значимости $\alpha = 0,05$

m	n					Дополнительные значения для $n=3$	
	3	4	5	6	7	m	S
3			64,4	103,9	157,3	9	54,0
4		49,5	88,4	143,3	217,0	12	71,9
5		62,6	112,3	182,4	276,2	14	83,8
6		75,7	136,1	221,4	335,2	16	95,8
8	48,1	101,7	183,7	299,0	453,1	18	107,7
10	60,0	127,8	231,2	376,7	571,0		
15	89,8	192,9	349,8	570,5	864,9		
20	119,7	258,0	468,5	764,4	1158,7		

ИСПОЛЬЗУЕМЫЕ В КНИГЕ ОБОЗНАЧЕНИЯ

Исходные наблюдения

n — число статистически обследованных объектов, объем выборки из многомерной генеральной совокупности;

p — число объясняющих (предикторных) переменных, регистрируемых на каждом из объектов;

$x_i^{(k)}$ — обозначение k-й объясняющей переменной на i-м обследуемом объекте;

y_i — значение исследуемого результирующего показателя («отклика») на i-м обследованном объекте;

$X_i = \begin{pmatrix} x_i^{(1)} \\ \vdots \\ x_i^{(p)} \end{pmatrix}$ — вектор-столбец значений p объясняющих переменных, зарегистрированных на i-м обследованном объекте;

$(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)}; y_i^{(1)}, y_i^{(2)}, \ldots, y_i^{(l)})$ — значения входных («объясняющих») $(x^{(k)})$ и выходных («результирующих») $(y^{(t)})$ переменных, зарегистрированные в i-м наблюдении (или на i-м обследованном объекте);

$\mathbf{B}_n = \{x_i^{(1)}, \ldots, x_i^{(p)}; y_i\}_{i=\overline{1, n}}$ — выборка объема n, исходные статистические данные;

$\mathbf{B}$ — система подвыборок выборки $\mathbf{B}_n$;

$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ — вектор-столбец наблюденных значений результирующей переменной («отклика»);

$$\mathbf{X} = \begin{pmatrix} \psi_0(X_1) & \psi_1(X_1) \ldots & \psi_m(X_1) \\ \psi_0(X_2) & \psi_1(X_2) \ldots & \psi_m(X_2) \\ \ldots & \ldots & \ldots \\ \psi_0(X_n) & \psi_1(X_n) \ldots & \psi_m(X_n) \end{pmatrix}$$

матрица (размера $n \times (m+1)$) исходных данных по объясняющим (предикторным) переменным, *матрица плана* (здесь $\psi_0(X)$, $\psi_1(X)$, ..., $\psi_m(X)$ — известные базисные функции, по которым разложена функция регрессии $f(X; \Theta)$); в частном случае *линейной* (по объясняющим переменным) модели регрессии матрица плана имеет вид:

$$\mathbf{X} \begin{pmatrix} 1 & x_1^{(1)} \ldots x_1^{(p)} \\ 1 & x_2^{(1)} \ldots x_2^{(p)} \\ \ldots & \ldots \\ 1 & x_n^{(1)} \ldots x_n^{(p)} \end{pmatrix}.$$

В гл. 3 $\mathbf{X}$ используется для обозначения матрицы исходных данных таблицы сопряженности (т. е. ее элемент x_{ij} — это число объектов в двумерной выборке объема n, отнесенных по первой случайной компоненте к градации i, а по второй случайной компоненте — к градации j), а в гл. 6 — для обозначения некоторого подмножества области определения исследуемой функции регрессии $f(X; \Theta)$.

Законы распределения вероятностей и их числовые характеристики

$N_k(M, \Sigma)$ — k-мерный нормальный закон распределения вероятностей с вектором (столбцом) средних значений M и ковариационной матрицей Σ (если контекст не требует уточнения размерности закона, нижний индекс может быть опущен);

$F(\nu_1, \nu_2; a)$ — нецентральное F-распределение с числами степеней свободы числителя и знаменателя соответственно ν_1 и ν_2 и с параметром нецентральности a;

$\xi \in N(M, \Sigma)$ — обозначение факта: «случайная величина ξ имеет распределение $N(M, \Sigma)$»;

u_q — квантиль уровня q(q-квантиль) стандартного нормального распределения;

$\chi^2_\alpha(\nu)$ — $100\alpha\%$-ная точка χ^2-распределения с ν степенями свободы;

$t_\alpha(\nu)$ — $100\alpha\%$-ная точка распределения Стьюдента с ν степенями свободы;

$v^2_\alpha(\nu_1, \nu_2)$ — $100\alpha\%$-ная точка центрального F-распределения с числами степеней свободы числителя и знаменателя соответственно ν_1 и ν_2;

p_{ij} — элемент последовательности чисел, задающей двумерный (дискретный) закон распределения вероятностей: вероятность, что случайно извлеченный объект будет отнесен по первой компоненте к i-й градации, а по второй — к j-й;

$$p_{i.} = \sum_j p_{ij};\quad p_{.j} = \sum_i p_{ij};$$

$\eta(X) = (\eta \mid \xi = X)$—случайная величина η, анализируемая при условии, что значение другой случайной величины (ξ) зафиксировано и равно X;

$\mathbf{E}\xi$ — теоретическое среднее (математическое ожидание) случайной величины ξ;

$\mathbf{D}\xi$ — дисперсия случайной величины ξ;

$\mathbf{E}(\eta \mid \xi = X)$ — условное среднее значение случайной величины η, вычисленное при условии, что значение другой случайной величины ξ зафиксировано на уровне X;

$\mathbf{D}(\eta \mid \xi = X)$ — условная дисперсия случайной величины η, вычисленная при условии, что значение другой случайной величины ξ зафиксировано на уровне X;

$\mathrm{cov}(\xi, \eta) = \mathbf{E}[(\xi - \mathbf{E}\xi)(\eta - \mathbf{E}\eta)]$ — ковариация случайных величин ξ и η;

Если $\xi = \begin{pmatrix} \xi^{(1)} \\ \vdots \\ \xi^{(p)} \end{pmatrix}$ и $\eta = \begin{pmatrix} \eta^{(1)} \\ \vdots \\ \eta^{(m)} \end{pmatrix}$, то

$\mathrm{cov}(\xi, \eta) = \mathbf{\Sigma}$ — ковариационная матрица (размера $(p+m) \times (p+m)$) вектора $(\xi', \eta')'$, причем она в ряде мест книги разбита на подматрицы по следующей схеме:

$$\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{\xi\xi} & \Sigma_{\xi\eta} \\ \Sigma_{\eta\xi} & \Sigma_{\eta\eta} \end{pmatrix},$$

где подматрицы $\Sigma_{\xi\xi}$, $\Sigma_{\xi\eta}$, $\Sigma_{\eta\xi}$ и $\Sigma_{\eta\eta}$ имеют соответственно размеры $p \times p$, $p \times m$, $m \times p$, $m \times m$.

Корреляционный анализ

$$r = \frac{\mathbf{E}[(\xi - \mathbf{E}\xi)(\eta - \mathbf{E}\eta)]}{[\mathbf{E}(\xi - \mathbf{E}\xi)^2 \cdot \mathbf{E}(\eta - \mathbf{E}\eta)^2]^{\frac{1}{2}}}$$ — *коэффициент коорреляции* между случайными величинами ξ и η;

$$\hat{r} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - y)^2\right]^{\frac{1}{2}}}$$ —

выборочный коэффициент корреляции между ξ и η (здесь x_i и y_i — i-е наблюденные значения случайных величин соответственно ξ и η, а $\bar{x} = \left(\sum_{i=1}^{n} x_i\right) \Big/ n$ и $\bar{y} = \left(\sum_{i=1}^{n} y_i / n\right)$;

$$I_{\eta \cdot \xi} = \sqrt{1 - \frac{\bar{\sigma}^2_{\eta(X)}}{\sigma^2_\eta}}$$ — *индекс корреляции*, характеризующий тесноту статистической связи между η и ξ в общем случае (здесь $\sigma^2_\eta = \mathbf{D}\eta$ — безусловная дисперсия случайной величины η, а $\bar{\sigma}^2_{\eta(X)}$ — усредненная по различным значениям X случайной величины ξ величина условной дисперсии $\mathbf{D}(\eta \mid \xi = X)$);

$$\hat{\rho}^2_{\eta\cdot\xi} = \frac{\sum_{i=1}^{k} m_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{k}\sum_{j=1}^{m_i}(y_{ij}-\bar{y})^2}$$ — *корреляционное отношение*, характеризующее тесноту статистической связи между η и ξ при разбиении диапазона изменения объясняющей (предикторной) переменной на k интервалов группирования (здесь y_{ij} — j-е наблюденное значение результирующей переменной η в i-м интервале группирования, $\bar{y}_i = \left(\sum_{j=1}^{m_i} y_{ij}\right)\Big/ m_i$ — средняя величина всех m_i наблюденных значений η, оказавшихся в i-м интервале группирования, а $\bar{y} = \left(\sum_{i=1}^{k} m_i\bar{y}_i\right)\Big/ n$ — общее среднее значение всех n наблюденных значений случайной величины η);

$I(\xi, \eta) = H(\xi) + H(\eta) - H(\xi, \eta)$ — информационная мера статистической связи между дискретными случайными величинами ξ и η, где $H(\zeta) = -\sum_z p_\zeta(z)\ln p_\zeta(z) = -\mathbf{E}(\ln p_\zeta(\zeta))$ — энтропия случайной величины ζ (здесь $p_\zeta(z) = P\{\zeta = z\}$, а суммирование производится по всем возможным значениям случайной величины ζ);

$r_{ij.X^{(i,j)}}$ — *частный (очищенный) коэффициент корреляции* между компонентами $x^{(i)}$ и $x^{(j)}$ вектора $X = (x^{(0)}, x^{(1)}, \ldots, x^{(p)})'$ (вычислен при условии, что все остальные компоненты $X^{(i,j)}$ вектора X зафиксированы на некотором постоянном уровне);

$r_{ij.(k_1 k_2 \ldots k_q)}$ — частный коэффициент корреляции между $x^{(i)}$ и $x^{(j)}$ при условии фиксации на постоянных уровнях компонент $x^{(k_1)}, x^{(k_2)}, \ldots, x^{k_q}$;

$R_{\eta\cdot\xi} = R_{\eta(\xi^{(1)}\ldots\xi^{(p)})} = \left(\sqrt{1-(1-r^2_{01})(1-r^2_{02(1)}) \times (1-r^2_{03(12)})\cdots(1-r^2_{0p(12\ldots p-1)})}\right)$ — множественный коэффициент корреляции между результирующей переменной $\eta = \xi^{(0)}$ и объясняющими переменными $\xi^{(1)}, \ldots, \xi^{(p)}$;

$R^2_{\eta\cdot\xi}$ — коэффициент детерминации между η и ξ;

$\hat{\tau}^{(S)}$ — ранговый коэффициент корреляции Спирмэна;

$\hat{\tau}^{(K)}$ — ранговый коэффициент корреляции Кендалла;

$W(m)$ — коэффициент конкордации (согласованности), измеряющий степень согласованности m различных ранжировок одних и тех же объектов.

Регрессионный, дисперсионный и ковариационный анализ

$f(X; \Theta) = \mathbf{E}(\eta \mid \xi = X)$ — *функция регрессии* результирующей переменной η по объясняющим переменным ξ (параметрическая запись):

$\rho(u)$ — *функция потерь*, измеряющая убытки от неточности восстановления значения $\eta(X) = \{\eta \mid \xi = X\}$ с помощью функции $f_a(X)$, где $u = u(X; f_a) = \eta(X) - f_a(X)$, а $f_a(X)$ — некоторая аппроксимация неизвестной функции регрессии $f(X)$;

$\Delta(f_a) = \mathbf{E}\rho\,(u(X;\ f_a))$ — *теоретический критерий адекватности* модели $f_a(X)$;

$\widehat{\Delta}_n\,(f_a) = \frac{1}{n}\sum_{i=1}^{n}\rho\,(u\,(X_i;\ f_a))$ — *выборочный критерий адекватности* модели $f_a(X)$;

$\mathbf{F}$ — *класс допустимых решений* (класс функций, в рамках которого подыскивается наилучшая аппроксимация для $f(X) = \mathbf{E}(\eta \mid \xi = X)$;

$f_a^{\Delta}(X) = \arg\min_{f_a \in \mathbf{F}} \Delta\,(f_a)$ — *функция Δ-регрессии*;

$\Theta = (\theta_0,\ \theta_1,\ \ldots,\ \theta_m)'$ — вектор-столбец неизвестных параметров, от которых зависит уравнение искомой функции регрессии $f(X;\ \Theta)$;

θ — статистическая оценка векторного параметра Θ;

$\psi(X) = (\psi_0(X),\ \psi_1(X),\ \ldots,\ \psi_m\,(X))'$ — вектор-столбец базисных функций $\psi_0(X),\ \ldots,\ \psi_m\,(X)$, по которым разложена функция регрессии $f(X;\ \Theta)$;

$f(X;\ \Theta) = \psi'(X)\cdot\Theta = \theta_0\psi_0(X) + \theta_1\psi_1(X) + \ldots + \theta_m\cdot\psi_m(X)$ — функция регрессии, разложенная в системе базисных функций $\psi(X)$, линейная по параметрам;

$\Sigma_{\hat{\Theta}}$ — ковариационная матрица оценок $\widehat{\Theta}$;

$M(\eta,\ X;\ \Theta) = 0$ —общая (неявная) запись модели регрессии η по X;

$X_{\text{д}} = (x_{\text{д}}^{(1)},\ldots,x_{\text{д}}^{(k)})$ — индикаторные переменные в схеме ковариационного анализа, соответствующие k возможным типам условий эксперимента (нижний индекс $_{\text{д}}$ показывает, что эти переменные относятся к «дисперсионной части» модели ковариационного анализа);

$\Theta_{\text{д}} = (\theta_{\text{д}1},\ \ldots,\ \theta_{\text{д}k})'$ — параметры (неизвестные) модели ковариационного анализа, определяющие сравнительный эффект влияния каждого из k типов условий эксперимента на исследуемый результирующий показатель.

СПИСОК ЛИТЕРАТУРЫ

1. Абрамовиц М., Стиган И. Справочник по специальным функциям (с формулами, графиками и математическими таблицами). Пер. с англ. — М.: Наука, 1979. — 831 с.
2. Адлер Ю. П. Введение в планирование эксперимента. — М.: Металлургия, 1969. — 159 с.
3. Адлер Ю. П. Предпланирование эксперимента. — М.: Знание, 1980. — 72 с.
4. Азарян О. В. Прогноз технико-экономических показателей проектируемых гидросооружений с помощью метода типологической регрессии. — В кн.: II Всесоюз. школа-семинар «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа» (сент., 1983 г.): Тез. докл. М., 1983, с. 135—137.
5. Азгальдов Г. Г. Теория и практика оценки качества товаров. — М.: Экономика, 1982. — 256 с.
6. Айвазян С. А. Ковариационный анализ. — В кн.: Математическая энциклопедия. М., 1979, т. 2, с. 901—902.
7. Айвазян С. А. Конфлюэнтный анализ. — В кн.: Математическая энциклопедия. М., 1979, т. 2, с. 1083.
8. Айвазян С. А. Многомерный статистический анализ. — В кн.: Математическая энциклопедия. М., 1982, т. 3, с. 731—738.
9. Айвазян С. А. Об опыте применения экспертно-статистического метода построения неизвестной целевой функции. — В кн.: Многомерный статистический анализ в социально-экономических исследованиях. М., 1974, с. 56—86.
10. Айвазян С. А. Статистическое исследование зависимостей. — М.: Металлургия, 1968. — 227 с.
11. Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. — М.: Статистика, 1974. — 240 с.
12. Айвазян С. А., Богдановский И. М. Методы статистического исследования парных зависимостей в схеме конфлюэнтного анализа и их применения. — Заводская лаборатория, 1974, т. 40, № 3, с. 285—295.
13. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. О структуре и содержании пакета программ по прикладному статистическому анализу. — В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980, с. 7—62.
14. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Основы моделирования и первичная обработка данных. — М.: Финансы и статистика, 1983. — 472 с.

15. Айвазян С. А., Розанов Ю. А. Некоторые замечания к асимптотически-эффективным линейным оценкам коэффициентов регрессии. — Труды Математического института им. В. А. Стеклова АН СССР, 1964, т. LXXI, с. 24—36.
16. Айвазян С. А., Тамарин А. А. Методика нахождения татировочных зависимостей для косвенного контроля технологических параметров изготовления железобетонных изделий и их конструкционных характеристик. — В кн.: Неразрушающие методы контроля качества железобетонных конструкций. М., 1972, с. 58—76.
17. Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание. Пер. с англ. — М.: Наука, 1977. — 223 с.
18. Анализ авторегрессий: Сб. статей. Пер. с англ. — М.: Статистика, 1978. — 231 с.
19. Алгоритмы многомерного статистического анализа и их применения. — М.: ротапринт ЦЭМИ, 1975. — 176 с.
20. Андерсон Т. Введение в многомерный статистический анализ. Пер. с англ. — М.: Физматгиз, 1963. — 500 с.
21. Андерсон Т. Статистический анализ временных рядов. Пер. с англ. — М.: Мир, 1976. — 755 с.
22. Апраушева Н. Н., Конаков В. Д. Использование непараметрических оценок в регрессионном анализе. — Заводская лаборатория, 1973, т. 39, № 5, с. 566—569.
23. Аптон Г. Анализ таблиц сопряженности. Пер. с англ. — М.: Финансы и статистика, 1982. — 143 с.
24. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. Пер. с англ. — М.: Мир, 1982. — 486 с.
25. Бард Й. Нелинейное оценивание параметров. Пер. с англ. — М.: Статистика, 1979. — 349 с.
26. Баруча-Рид А. Т. Элементы теории марковских процессов и их приложения. Пер. с англ. — М.: Наука, 1969. — 511 с.
27. Бешелев С. Д., Гурвич Ф. Г. Математико-статистические методы экспертных оценок. — М.: Статистика, 1980. — 263 с.
28. Бокс Дж., Дженкинс Г. Анализ временных рядов: Прогноз и управление. — М.: Мир, 1974. Вып. 1. — 406 с.; вып. 2. — 224 с.
29. Болч Б., Хуань К. Дж. Многомерные статистические методы для экономики. Пер. с англ. — М.: Статистика, 1979. — 317 с.
30. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. — М.: Наука, 1965. — 464 с.
31. Браун М. Теория и измерение технического прогресса. Пер. с англ. — М.: Статистика, 1971. — 208 с.
32. Бухштабер В. М., Маслов К. В., Маркин В. Г. Обратные задачи прикладной статистики и томография. — В кн.: II Всесоюз. школа-семинар «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа» (сент. 1983 г.): Тез. докл. М., 1983, с. 26—33.
33. Ван-дер-Варден Б. Л. Математическая статистика. Пер. с нем. — М.: ИЛ, 1960. — 434 с.
34. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979. — 447 с.
35. Васильев Ф. П. Численные методы решения экстремальных задач. — М.: Наука, 1980. — 520 с.
36. Веревка О. В. Функциональные возможности пакета Δ-СТАТ. — В кн.: II Всесоюз. школа-семинар «Программно-ал-

горитмическое обеспечение прикладного многомерного статистического анализа» (сент. 1983 г.): Тез. докл. М., 1983, с. 291—293.
37. Верексков А. И. Сходимость некоторых алгоритмов минимизации, не использующих производных. — В кн.: Методы исследования сложных систем: Тр. семинара аспирантов и молодых специалистов ВНИИ системных исследований. М., 1981, с. 58—62.
38. Верескoв А. И., Левин В. Е., Федоров В. В. Регуляризированный м. н. к. без производных. — В кн.: Линейная и нелинейная параметризация в задачах планирования экспериментов. М., 1981, с. 20—27.
39. Воеводин Г. В. Вычислительные основы линейной алгебры. — М.: Наука, 1977. — 304 с.
40. Гаврилец Ю. Н. Социально-экономическое планирование: Системы и модели. — М.: Экономика, 1974. — 174 с.
41. Гренджер К., Хатанака М. Спектральный анализ временных рядов в экономике. — М.: Статистика, 1972. — 312 с.
42. Деев А. Д., Бурлуцкий Г. И. О распределении условной невязки прогноза в модели множественной регрессии и отборе информативных признаков. — В кн.: II Всесоюз. науч.-техн. конференция «Применение многомерного статистического анализа в экономике и оценке качества продукции»: Тез. докл. Тарту, 1981, с. 208—213.
43. Демиденко Е. З. Линейная и нелинейная регрессии. — М.: Финансы и статистика, 1981. — 302 с.
44. Демиденко Е. З. Гребневая регрессия. — М., ИМЭМО АН СССР, 1982 (препринт). — 126 с.
45. Денисов В. И. Математическое обеспечение системы «экспериментатор». — М.: Наука, 1977. — 132 с.
46. Джонстон Дж. Эконометрические методы. Пер. с англ. — М.: Статистика, 1980. — 446 с.
47. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Пер. с англ. — М.: Статистика, 1973. — 392 с.
48. Дуда Р., Харт П. Распознавание образов и анализ сцен. Пер. с англ. — М.: Мир, 1976. — 511 с.
49. Дукарский О. М., Левит Б. Я. Некоторые применения непараметрических оценок регрессии. — В кн.: Многомерный статистический анализ в социально-экономических исследованиях. М., 1974, с. 31—37.
50. Езекиэл М., Фокс К. А. Методы анализа корреляций и регрессий. Пер. с англ. — М.: Статистика, 1966. — 559 с.
51. Енюков И. С. Методы оцифровки неколичественных признаков. — В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980, с. 309—316.
52. Енюков И. С. Оценивание параметров и критерии отбора информативных переменных в линейных регрессионных моделях со случайными аргументами. — В кн.: II Всесоюз. науч.-техн. конференция «Применение многомерного статистического анализа в экономике и оценке качества продукции»: Тез. докл. Тарту, 1981, с. 214—218.
53. Епишин Ю. Г. Об оценках параметров регрессии по методу наименьших абсолютных отклонений. — Экономика и математические методы, 1974, т. 10, вып. 5, с. 1023—1028.
54. Завьялов Ю. С., Квасов Б. И., Мирошниченко В. Л. Методы сплайн-функций. — М.: Наука, 1980. — 352
55. Загоруйко Н. Г. Эмпирическое предсказание. — Новосибирск: Наука, 1979. — 124 с.

56. З а й ц е в а Л. М. Структурный подход к определению взаимосвязей в системе случайных величин. — Изв. АН СССР. Техническая кибернетика, 1984, № 4, с. 61—82.
57. З а к с Л. Теория статистических выводов. Пер. с англ. — М.: Мир, 1975. — 776 с.
58. З а р у ц к и й В. И. Классификация нормальных векторов простой структуры в пространстве большой размерности. — В кн.: Прикладной многомерный статистический анализ. М., 1978, с. 37—51.
59. З а р у ц к и й В. И. О выделении некоторых графов связей для нормальных векторов в пространстве большой размерности. — В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980, с. 189—208.
60. З е л ь н е р А. Байесовские методы в эконометрии. Пер. с англ.— М.: Статистика, 1980. — 439 с.
61. К а р а п е т я н К. А. Об одном статистическом критерии проверки гипотезы о структуре многомерных наблюдений. — В кн.: Многомерный статистический анализ в социально-экономических исследованиях. М., 1974, с. 294—308.
62. К а ц М. Вероятность и смежные вопросы в физике. — М.: Мир, 1975. — 406 с.
63. К е м е н и Дж., С н е л л Дж. Кибернетическое моделирование: Некоторые приложения. Пер. с англ. — М., Советское радио, 1972. — 192 с.
64. К е н д а л л М. Дж., С т ь ю а р т А. Теория распределений. Пер. с англ. — М.: Наука, 1966. — 587 с.
65. К е н д а л л М. Дж., С т ь ю а р т А. Статистические выводы и связи. Пер. с англ. — М.: Наука, 1973. — 899 с.
66. К е н д а л л М. Дж., С т ь ю а р т А. Многомерный статистический анализ и временные ряды. Пер. с англ. — М.: Наука, 1976. — 736 с.
67. К е н д э л М. Ранговые корреляции. — М.: Статистика, 1975. — 214 с.
68. К и с е л е в Н. И. Экспертно-статистический метод определения функции предпочтения по результатам парных сравнений. — В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980, с. 111—123.
69. К о н а к о в В. Д. О структуре и содержании библиотеки программ по разделу «Статистическое исследование зависимостей». —В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа, М., 1980, с. 63—92.
70. К о к с Д., Х и н к л и Д. Теоретическая статистика. Пер. с англ. — М.: Мир, 1978. — 560 с.
71. К р а м е р Г Математические методы статистики. — 2-е изд. Пер. с англ. — М.: Мир, 1975. — 648 с.
72. К у к с Я. П., В и й к м а н н Э. В., П у к к К. Г. Программы корреляционного и одномерного регрессионного анализа.—Таллин: АН ЭССР, 1980, ротапринт. — 50 с.
73. К у к с Я. П., Т и й т с Т. В., В и й к м а н н Э. В. Программы множественного регрессионного анализа. — Таллин: АН ЭССР, 1979. — 61 с.
74. К у к с Я. П., Т и й т с Т. В. Программы дисперсионного анализа.— Таллин: АН ЭССР, 1980, ротапринт. — 72 с.
75. К у л ь б а к С. Теория информации и статистика. Пер. с англ. — М.: Наука, 1967. — 408 с.

76. Лбов Г. С. Методы обработки разнотипных экспериментальных данных. — Новосибирск: Наука, 1981. — 180 с.
77. Линник Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. — М.: Физматгиз, 1958. — 333 с.
78. Лоэв М. Теория вероятностей Пер. с англ. — М.: ИЛ, 1962. — 719 с.
79. Малолеткин Г. Н., Мельников И. Н., Хашин В. М. Об алгоритмах выбора наилучшего подмножества признаков в регрессионном анализе. — В кн.: Вопросы кибернетики: Теоретические проблемы планирования эксперимента. М., 1977, вып. 35, с. 110—145.
80. Маленво Э. Статистические методы в эконометрии. Пер. с франц.— М.: Статистика, 1975, вып. 1, 422 с.; 1976, вып. 2, 325 с.
81. Малютов М. Б., Мартинес-Кресно К. Х. Оценивание тренда и компонент дисперсии линейной регрессионной модели. — В кн.: Математическая статистика. Пермь, 1980, с. 114—134.
82. Маракуев А. В. Статистические модели для формирования нормативов в судоремонтных работах. — В кн.: II Всесоюз. науч.-техн. конференция «Применение многомерного статистического анализа в экономике и оценке качества продукции»: Тез. докл. Тарту, 1981, с. 123—127.
83. Математические алгоритмы и программы для малых ЭВМ. — М.: Финансы и статистика, 1981, с. 123—127.
84. Математическое обеспечение ЕС ЭВМ. — Минск: Институт математики АН БССР, 1980, вып. 1, с. 202.
85. Матюха И. Я. Статистика бюджетов населения. — М.: Статистика, 1967. — 248 с.
86. Мацевич Д. А., Петрович М. Л., Федоров В. В. Экспериментальное сравнение методов оценивания параметров регрессионных моделей в случае ошибок в независимых переменных. — В кн.: Программное обеспечение ЭВМ, Минск, 1982, вып. 35, с. 42—56.
87. Мельников Н. Н., Малолеткин Г. Н. Стандартное математическое обеспечение ЕС ЭВМ для анализа регрессионных экспериментов. — В кн.: II Всесоюз. научн.-техн. конференция «Применение многомерного статистического анализа в экономике и оценке качества продукции»: Тез. докл. Тарту, 1981, с. 379—382.
88. Мешалкин Л. Д. Использование весовой функции при оценке регрессионной зависимости. — В кн.: Многомерный статистический анализ в социально-экономических исследованиях. М., 1974, с. 25—30.
89. Мешалкин Л. Д., Курочкина А. И. Новый подход к параметризации регрессионных зависимостей. — В кн.: Исследования по математической статистике: Записки научных семинаров ЛОМИ АН СССР. Л., 1979, т. 87, с. 79—86.
90. Мешалкин Л. Д., Струнин Б. М. Методика составления таблиц перевода чисел твердости. — Заводская лаборатория, 1967, т. 33, № 11, с. 1408—1417.
91. Мирзоев А. А. Система программ логлинейного анализа социологической информации. — В кн.: I Всесоюз. школа-семинар «Программно-алгоритмическое обеспечение прикладного статистического анализа»: Тез. докл. Ереван, 1979, с. 259—260.
92. Миркин Б. Г. Анализ качественных признаков и структур. — М.: Статистика, 1980. — 319 с.

93. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Пер. с англ. — М.: Финансы и статистика, 1982, вып. 1. — 224 с.; вып. 2. — 240 с.
94. Мудров В. И., Кушко В. Л. Методы обработки измерений. — М.: Советское радио, 1976. — 192 с.
95. Налимов В. В. Применение математической статистики при анализе вещества. — М.: Физматгиз, 1969. — 340 с.
96. Нейлор Т. Машинные имитационные эксперименты с моделями экономических систем. Пер. с англ. — М.: Мир, 1975. — 500 с.
97. Орлов А. И. Оценка размерности модели регрессии. — В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980, с. 92—99.
98. ОТЭКС: Пакет прикладных программ для обработки таблиц экспериментальных данных (версия 3.0). Новосибирск, 1981. — 27 с.
99. Пакет программ по прикладному статистическому анализу (ППСА). — М.: ЦЭМИ АН СССР, 1983. — 187 с.
100. Пакет прикладных программ статистической обработки данных на ЕС ЭВМ. — Киев: Ин-т кибернетики АН УССР, 1979. — 66 с.
101. Пакет промышленных программ для анализа и прогноза временных рядов/Френкель А. Д., Гарбер Е. В., Шифрин Г. М., Горелик Н. А. — В кн.: II Всесоюз. школа-семинар «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа»: Тез. докл., М., 1983, с. 167—169.
102. Парлетт П. Б. Симметричная проблема собственных значений: Численные методы. — М.: Мир, 1983. — 382 с.
103. Петрович М. Л. Регрессионный анализ и его математическое обеспечение на ЕС ЭВМ. — М.: Финансы и статистика, 1982. — 199 с.
104. Пикялис В., Раудис Ш. Общее описание пакета СОРРА-2. Входной язык. Условия применения. — В кн.: Статистические проблемы управления. — Вильнюс: ИМК АН ЛитССР, 1982, вып. 58, с. 9—26.
105. Пинкава Я. Вероятностные распределения в задачах статистического ранжирования. — Вопросы кибернетики. Экспертные оценки. — М.: АН СССР, 1979, с. 34—52.
106. Пирогов Г. Г., Федоровский Ю. П. Проблемы структурного оценивания в эконометрии. — М.: Статистика, 1979. — 328 с.
107. Поляк Б. Т. Методы минимизации функций многих переменных. — Экономика и математические методы, 1967, т. 3, № 6, с. 881—902.
108. Поляк Б. Т. Метод сопряженных градиентов в задачах на экстремум. — Журн. вычисл. матем. и матем. физ., 1969, т. 9, № 4, с. 807—821.
109. Поляк Б. Т. Сходимость методов возможных направлений в экстремальных задачах. — Журн. вычисл. матем. и матем. физ., 1971, т 11, № 4, с. 855—869.
110. Поляк Б. Т., Цыпкин Я. З. Адаптивные алгоритмы оценивания (сходимость, оптимальность, стабильность). — Автоматика и телемеханика, 1979, № 3, с. 71—84.
111. Программное обеспечение ЭВМ. — Минск: ИМ АН СССР, 1982, вып. 36. — 72 с.
112 Программный комплекс по планированию эксперимента для ЕС ЭВМ/Бродский Л. И., Малолеткин Г. Н., Мельников Н. И. — В кн.: I Всесоюз. школа-семинар «Программно-алгоритмическое

обеспечение прикладного статистического анализа»: Тез. докл. Ереван. 1979, с. 140—146.
113. Прохорскас Р. П., Жюжнис В. Е., Мисюнене Н. Б. Применение некоторых классификаторов для прогнозирования отдаленных итогов инфаркта миокарда. — В кн.: Проблемы ишемической болезни сердца. — Вильнюс, 1976, с. 261—267.
114. Пуарье Д. Эконометрия структурных изменений (с применением сплайн-функций). Пер. с англ. — М.: Финансы и статистика, 1981. — 184 с.
115. Пшеничный Б. Н., Данилин Ю. М. Численные методы в экстремальных задачах. — М.: Наука, 1975. — 319 с.
116. Райбман Н. С., Дорофеюк А. А., Касавин А. Д. Идентификация технологических объектов методами кусочной аппроксимации. — М.: Ин-т проблем управления, 1977. — 70 с.
117. Рао С. Р. Линейные статистические методы и их применения. Пер. с англ. — М.: Наука, 1968. — 548 с.
118. Сборник научных программ на Фортране. — М.: Статистика, 1974, вып. 1. — 216 с.
119. Себер Дж. Линейный регрессионный анализ. Пер. с англ. — М.: Мир, 1980. — 456 с.
120. Сильвестров Д. С., Клесов О. И., Ремнев В. Н. Пакеты прикладных программ по анализу временных рядов ПАРИС и МАВР. — В кн.: II Всесоюз. школа-семинар «Программно-алгоритмическое обеспечение многомерного статистического анализа» (сент. 1983 г.): Тез. докл. М., 1983, 165—166.
121. Система обработки разнотипных данных SITO: Интерактивный вариант/В. В. Александров, А. И. Алексеев, Н. Д. Горский, А. М. Никифоров. — Л., ЛНИВЦ АН СССР, 1982, препринт. — 45 с.
122. Система статистического анализа и обработки наблюдений на ЕС ЭВМ/О. М. Дукарский, Н. М. Кисурина, Ю. А. Кошевник и др. — Энергетическое строительство, 1979, № 10, с. 69—72.
123. Смоляк С. А. Оптимальное восстановление функций и связанные с ним геометрические характеристики множеств. — В кн.: Тр. 3 зимней школы по математическому программированию и смежным вопросам. — М.: ЦЭМИ АН СССР, 1970, вып. 3, с. 509 — 557.
124. Смоляк С. А., Титаренко Б. П. Устойчивые методы оценивания. — М.: Статистика, 1980. — 208 с.
125. Степнов Н. Н. Линейный регрессионный анализ результатов усталостных испытаний алюминиевых сплавов. — Заводская лаборатория, 1963, т. 29, № 10, с. 1212—1214.
126. Тейл Г. Эконометрические прогнозы и принятие решений. Пер. с англ. — М.: Статистика, 1971. — 488 с.
127. Тийт Э. М. Актуальные проблемы анализа данных и некоторые возможности их решения. — В кн.: II Всесоюз. науч.-техн. конференция «Применение многомерного статистического анализа в экономике и оценке качества продукции»: Тез. докл. Тарту, 1981, с. 64—79.
128. Типология потребления/Под ред. С. А. Айвазяна и Н. М. Римашевской. — М.: Наука, 1978. — 175 с.
129. Тоодинг Л. М. Система статистической обработки данных в вычислительном центре ТГУ: Тр. вычисл. центра, Тартуск. гос. ун-т, 1977, вып. 40, с. 3—7.
130. Тутубалин В. Н. Теория вероятностей. — М.: МГУ, 1972. — 68 с

131. Тюрин Ю. Н., Яхья А. Ю. Доверительное оценивание порядков на основе рангов. — Вопросы кибернетики. Экспертные оценки. М., 1979, с. 66—72.
132. Уилкинсон Дж., Райнис С. Справочник алгоритмов на языке Алгол: Линейная алгебра. Пер. с англ. — М.: Машиностроение, 1977. — 389 с.
133. Уилкинсон Дж. Алгебраическая проблема собственных значений. Пер. с англ.— М.: Наука, 1970.— 564 с.
134. Уилсон Р. Введение в теорию графов. Пер. с англ.— М.: Мир, 1977.— 207 с.
135. Успенский А. Б., Федоров В. В. Вычислительные аспекты метода наименьших квадратов при анализе и планировании регрессионных экспериментов.— М.: МГУ, 1975.— 168 с.
136. Федоров В. В. Теория оптимального эксперимента. — М.: Наука, 1971.— 197 с.
137. Федоров В. В. Оценивание параметров регрессии в случае вектор-наблюдения.— В кн.: Регрессионные эксперименты. М., 1977, с. 112—122.
138. Федоров В. В. Регрессионный анализ при наличии погрешностей в определении предиктора.— Вопросы кибернетики, 1978, вып. 47, с. 69—75.
139. Федотов А. М. DIAS — автоматизированная система хранения и статистической обработки экспериментальных данных (краткое описание). — Новосибирск: ВЦ СО АН СССР, 1978, ч. 1, препринт.— 36 с.
140. Феллер В. Введение в теорию вероятностей и ее приложения. Т.2. Пер. с англ.— М.: Мир, 1967.— 752 с.
141. Фишер Ф. Проблема идентификации в эконометрии. Пер. с англ.— М.: Статистика, 1978.— 224 с.
142. Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений. Пер. с англ.— М.: Мир, 1980.—280 с.
143. Форсайт Дж., Молер К. Численное решение систем линейных алгебраических уравнений. Пер. с англ. — М.: Мир, 1969.— 167 с.
144. Френкель А. А. Математические методы анализа динамики и прогнозирования производительности труда.— М.: Экономика, 1972.—190 с.
145. Химмельблау Д. Анализ процессов статистическими методами. Пер. с англ.— М.: Мир, 1973.— 957 с.
146. Химмельблау Д. Прикладное нелинейное программирование. Пер. с англ.— М.: Мир, 1975.— 534 с.
147. Ченцов Н. Н. Статистические решающие правила и опимальные выводы.— М.: Наука, 1972.— 520 с.
148. Шеффе Г. Дисперсионный анализ. Пер. с англ.—М.: Физматгиз, 1963.— 626 с.
149. Шурыгин А. М. Математические модели статистического оценивания.— В кн.: II Всесоюз. школа-семинар «Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа» (сент., 1983 г.): Тез. докл. М., 1983, с. 73—81.
150. Юдин А. Д. Об одной задаче оптимальной обработки результатов наблюдений.—В кн.: Алгоритмическое и программное обеспечение прикладного статистического анализа. М., 1980, с. 279—286.
151. Яковлев А. А., Ставицкая Н. А. Алгоритм выбора субоптимальной совокупности предикторов для множественной многомерной регрессии.—В кн.: Вопросы кибернетики. Нетрадицион-

ные подходы к планированию эксперимента.— М.: ВИНИТИ, 1981, с. 110—118.
152. Я н ч Э. Прогнозирование научно-технического прогресса. Пер. с англ.— М.: Прогресс, 1970.— 568 с.
153. A l v e y N., G a l w e y N., L a n e P. An introduction to GENSTAT.— London, N. Y.: Academic Press, 1982.— 152 p.
154. A n d e r s e n E. Discrete statistical models with social science applications. — North Holland, 1980.— 220 p.
155. A n d e r s o n T. W. On asymptotic distribution of estimate of parameter of stochastic difference equations.— Ann. Math. Statist. 1959, vol. 30, № 3, p. 676—687.
156. A n d r e w s D. F. A robust method for multiple linear regression.— Technometrics, 1974, vol. 16, № 4, p. 523—531.
157. B a c h a c o u J., M i l l i e r C., M a s s o n J. P. Manuel de la programmthèqe statistique AMANCE 81.— Verasilles: I. N. R. A., 1981, 516 p.
158. B a k e r R. J., R i c h a r d s o n M. G. GLIM-3.— In: COMPSTAT-82, Proc. in Comput. Statist.: Package and Facilities Present at COMPSTAT-82, Wien, 1982, p. 59—60.
159. B a r d Y. Comparison of Gradient Methods for the Solution of Non-linear Parametric Estimation Problems. — SIAM J. Numerical Analysis, 1970, vol. 7, p. 157—186.
160. B a r r A. J. et al. A User's Guide to SAS-76.— SAS Institute Inc., 1976.— 329 p.
161. B e a l e E. M. L. Confidence regions in nonlinear estimation. — J. Roy. Statist. Soc., 22, 1960, ser. B, p. 41—76.
162. B e a l e E. M. L. The scope of Jordan elimination in statistical computing.— J. Inst. Math. Appl., 1974, vol. 10, p. 138—140.
163. B e l s l e y D. A., K u h E., W e l s c h R. E. Regression diagnostics: Identifying influential data and sources of collinearity.— N. Y. etc.: John Wiley and sons, 1980.— 292 p.
164. B e n d e l R. B., A f i f i A. A. Comparison of Stopping Rules in Forward «Stepwise» Regression. — J. Amer. Statist. Assoc., 1977, vol. 72, p. 46—53.
165. B e n z é c r i J. P. et al. L'analyse des données. I La Taxonomie.— Paris. Dunod, 1973. — 611 p.
166. B e n z é c r i J. P. et al. L'analyse des données: II L'analyse des correspondances.— Dunod, 4-e ed., 1982.— 632 p.
167. B e r k s o n J. Are there two regression?— J. Amer. Statist. Assoc., 1950, vol. 45, p. 164—180.
168. B i s h o p Y. M., F i e n b e r g S. E., H o l l a n d P. W. Discrete Multivariate Analysis: Theory and Practice. — Cambridge: NIT — Press, 1975.— 607 p.
169. BMDP Biomedical Computer Programs. Ed. W. J. Dixon.— Univ. of California Press, 1979.— 880 p.
170. B o x G. E. P., C o x D. R. An analysis of transformations. J. Roy. Statist. Soc., ser. B, 1964, vol. 26, p. 211—243.
171. B r a d u D., G a b r i e l K. R. The biplot as a diagnostic tool for models of two-way tables.— Technometrics, 1978, vol. 20, p. 47—68.
172. C h a m b e r s J. Fitting nonlinear models: numerical techniques.— Biometrika, 1973, vol. 60, p. 1—13.
173. C h a n T. F., G o l u b G. H. Le V e q u e R. J. Updating Formulae and a Pairwise Algorithm for Computing Sample Variances. — In: Compstat-82, Proc. in Computational Statistics, 5 th Sympos. Wien, 1982, p. 30—41.

174. C h o w C. K. Tree dependence in normal distributions.— In: The 1970 International Symposium on Information Theory. The Netherlands, 1970, p. 2.9.
175. C h o w C. K., L i u C. N. Approximating discrete probability distributions with dependence trees.— IEEE Trans. Inform. Theory IT — 14, 1968, p. 462—467.
176. C h o w C. K., L i u C. N. An approach to structure adaptation in pattern recognition.— IEEE Trans. Sys. Sci. Cyb. SSC-2, 1966, p. 73—80.
177. C o h e n A. All admissible linear estimates of the mean vector.— Ann. of Math. Statistic., 1966, vol. 37, p. 458—463.
178. C o o k e D., G r a v e n A. H., C l a r k e G. M. Basic Statistical Computing.— London: Edward Arnold Ltd., 1982.— 156 p.
179. D e m p s t e r A. Covariance selection.— Biometrics, 1972, vol. 28, № 1, p. 167—175.
180. E f r o i m s o n M. A. Multiple regression analysis.— In: Mathematical Methods for Digital Computers, Ed. by Ralston A. and Wilf H. S., N. Y., 1960, p. 191—203.
181. F a r e b r o t h e r R. W. The minimum mean square error linear estimator and ridge regression.— Technometrics, 1975, vol. 17, № 1, p. 127—128.
182. F e d o r o v V. V. Regression problems with controllable variables, subject to error.— Biometrika, 1974, vol. 61, p. 49—56.
183. F i s h e r R. A. Statistical methods for Research workers, 10 th ed.— London: Oliver and Boyd, 1948.— 372 p.
184. F i s k P. R. Models of second kind in regression analysis. J. Roy. Statist. Soc. ser. B, 1967, vol. 29, p. 266—281.
185. F l e t c h e r R. Function Minimization without Evaluating Derivatives.— a Review, Comput. J., 1965, vol. 8, p. 33—41.
186. F l e t c h e r R., G r a n t J. A., H e b l e n H. D. The calculation of linear best L_p — approximations — a Review, Comput J., 1971, vol. 14, № 3, p. 276—279.
187. F o r s y t h e A. B. et al. A stoping rule for variable selection in multiple regression. — J. Amer. Statist. Assoc., 1973, vol. 68, № 341, p. 75—77.
188. F r a n s i s I. A survey of statistical Software.— Computational statist. and Data analysis, 1983, vol. 1, № 1, p. 17—27.
189. F u r n i v a l G. M. All possible regression with less computation.— Technometrics, 1971, vol. 13, p. 403—408.
190. F u r n i v a l G. M., W i l s o n R. W. M., Jr. Regressions by leaps and bounds.— Technometrics, 1974, vol. 16, p. 499—511.
191. G a r s i d e M. J. Some computational procedures for the best subset problem.— Appl. Statist., 1971, vol. 20, p. 8—15.
192. G o l d s t e i n M., Smith A. F. M. Ridge-type estimators for regression analysis.— J. Roy. Statist. Soc., 1974, Ser. B., vol. 36, p. 284—291.
193. G o l u b G. Numerical Methods for Solving Linear Least Squares Problems.— Numer. Math., 1965, vol. 7, p. 206—216.
194. G o l u b G., K a h a n W. Calculating the Singular Values and Pseudoinverse of a Matrix.— SIAM. J. Numer. Anal., 1965, Ser. B, vol. 2, p. 205—224.
195. G r e e n b e r g E. Minimum variance properties of principal component regression. J. Amer. Statist. Assoc., 1975, vol. 70, p. 194—197.
196. G r e n a n d e r U., R o s e n b l a t t M. Statistical analysis of stationary time series. — N. Y.: Wiley, 1966.— 300 p.

197. G r e v i l l e T. N. E. Theory and applications of spline functions — N. Y.: Academic Press, 1969. p. 20.
198. G u t t m a n L. The quantification of a class of attributes: a theory and method of scale construction.— In: The Prediction of Personal Adjustment. — Bulletin № 48 N. Y.: Social Science Research. Concil, 1941, p. 319—348.
199. H a b e r m a n S. J. Analysis of frequency data.— Chicago: Univ. of Chicago Press, 1974.— 208 p.
200. H a r t l e y H. O. Modified Gauss-Newton method for the fitting of nonlinear regression function.— Technometrics, 1961, vol. 3, p. 269—275.
201. H a w k i n s D. M. On the investigation of alternative regressions by principal component analysis.— Appl. Statist., 1973, vol. 22, № 3, p. 275—286.
202. H a w k i n s D. M. Relations between ridge regression and eigenanalysis of the augmented correlation matrix.— Technometrics, 1975, vol. 17, № 4, p. 477—480.
203. H i r s c h f e l d H. O. A connection between correlation and contingency.— Proceedings of Cambridge Philosophical Society, 1935, 31, p. 520—524.
204. H o c k i n g R. R. Griteria for selection of a subset regression: whith one should be used? — Technometrics, 1972, vol. 14, p. 967—970.
205. H o c k i n g R. R. The analysis and selection of variables in linear regression.— Biometrics, 1976, vol. 32, № 1, p. 1—49.
206. H o c k i n g R. R., L e s l i e R. N. Selection of the best subset in regression analysis.— Technometrics, 1967, vol. 9, p. 531—540.
207. H o c k i n g R. R., S p e e d F. M., L y n n M. J. A class of biased estimators in linear regression.— Technometrics, 1976, vol. 18, № 4, p. 425—438.
208. H o e r l A. E., K e n n a r d R. W. Ridge-regression. Biased estimation for non-ortogonal problems.— Technometrics, 1970, vol. 12, № 1, p. 55—68.
209. H o e r l A. E., K e n n a r d R. W. Ridge-regression. Applications to non-ortogonal problems.— Technometrics, 1970, vol. 12, № 1, p. 69—82.
210. H o r s t P. Measuring complex attitudes. J. of Social. Psychology, 1935, vol. 6, p. 369—374.
211. H u a n g N. Y., Unified Approach to Quadratically Convergent Algorithms for Function: Minimization.— J. Optim. Theory Applic., 1970, vol. 5, 6, p. 405—423.
212. H u b e r P. J. Robust Statistics: a review. Ann. of Math. Statist., 1972, vol. 43, p. 1041—1061.
213. H u b e r P. J. Robust estimation of a location parameter.— Ann. Math. Statist., 1964, vol. 35, p. 73—101.
214. H u b e r P. J. Robust regression: asyptotic, conjectures and Monte-Carlo.— Amer. Statist., 1973, vol. 1, № 5, p. 799—821.
215. J a e c k e l L. B. Robust estimates of location: symmetry and asummetric contamination.— Ann. Math. Statist., 1971, vol. 42, № 4, p. 1020—1034.
216. J a m e s W., S t e i n C. Estimation with guadratic loss. — In: Proc. Fourth Berkeley Symp. Math. Statist. and Prob, 1961, vol. 1, p. 361—379.
217. J o h n s o n D. E., G r a y b i l l F. A. An analysis of a two-way model with interaction and no replication. — J. Amer. Statist Assoc., 1972, vol. 67, p. 388—394.

218. IMSL Library Information, Fortran subroutines.— USA, IMSL Inc., 1981, p. 25.
219. Kendall M. G. A course in multivariate analysis. — London: Griffin, 1957.— 185 p.
220. La Motte L. R., Hocking R. R. Computational efficiency in the selection of regression variables.— Technometrics, 1970, vol. 12, p. 83—93.
221. Lebart L., Morineau A. SPAD — Systeme portable pour l' analyse des donnees.— Paris: Cesia, 1982.—243 p.
222. Leeuw J. Canonical analysis of Categorical Data.— The Netherlands Psychological Institute, Univ. of Leiden, 1973. — 120 p.
223. Lindley D. Regression lines and the linear functional relationship.— J. Roy. Statist. Soc., 1947, vol. 9, Suppl., p. 218—225.
224. Lingoes J. C. Geometric Representations of Relational Data: Readings in Multidimensional Scaling Ann. Arbor: Mathesis Press, 1977.— 165 p.
225. Mallows C. L. Some Comments on C_p.— Technometrics, 1973, vol. 15, № 4, p. 661—676.
226. Mantel N. Why stepdown procedures in variable selection? — Technometrics, 1970, vol. 12, № 3, p. 621—625.
227. Marquardt D. W. Generalised inverses, ridge regression, biased linear estimation and nonlinear estimation. — Technometrics, 1970, vol. 12, № 3, p. 591—612.
228. Mason R. L., Gunst R. F., Webster J. T. Regression analysis and problems of multicollinearity. — Communications in Statist., 1973, № 4, p. 277—292.
229. Massy W. F. Principal components regression in exploratory statistical research.— J. Amer. Statist. Assoc., 1965, vol. 60, № 2, p. 234—256.
230. Maung K. Measurement of association in contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children.— Ann. of Eugenics, 1941, vol. 11, p. 189—223.
231. Mayer L. S., Willke T. A. On biased estimation in linear models.— Technometrics, 1973, vol. 15, p. 497—508.
232. Nishisato Sh. Analysis of categorical data: duel scaling and its applications.— Toronto — Buffalo — London: Univ. of Toronto press, 1980.— 285 p.
233. Olkin I., Prat J. W. A Biased Estimation of Certain Correlation Coefficients.—Ann. Math. Statist., 1958, vol. 29, p. 201—211.
234. Pearson E. S. The analysis of variance in a cases of nonnormal variation.— Biometrika, 1931, vol. 23, p. 114—133.
235. Pearson K. On lines and planes closest fit to systems of points in space. — Phil. Mag., 1901, S. 6, vol. 2, p. 559—572.
236. Peckham G. A new method for minimizating a sum of squares without calculating gradients.— Computer J., 1970, vol. 13, p. 418—420.
237. Pereyra V. Iterative Methods for Solving Nonlinear Least Squares Problems.— SIAM J. Numerical Analysis, 1967, vol. 4, № 1, p. 27—36.
238. Plackett R. L. The Analysis of Categorical Data.— London: Griffin, 1974.— 159 p.
239. Plaut H. C. Ber. Fachausschüsse Dtsch. Glastechn. Ges., 1931, № 19, p. 121—130.
240. Powell M. I. D. A Survey of Numerical Methods for Unconstrained Optimization. SIAM Rev., 1970, vol. 12, № 1, p. 79—97.

241. P-STAT: Conversational Statistical and Data Management Software.— In: COMPSTAT-82, Proc. in Computational Statist., 5-th Symp., Packages and Facilities Present at COMPSTAT-82, Wien, 1982, p. 47—48.
242. Ralston M. L., Jennrich R. I. Dud, a derivative-free algorithm for non-linear estimation.— Technometrics, 1978, vol. 20, p. 7—14.
243. Ramsay J. O. A comparative study of several robust estimates of slope, intersept and scale in linear regression. J. Amer. Statist. Assoc., 1977, vol. 72, № 3, p. 608—615.
244. Rao C. R. Simultaneous estimation of parameters in different linear models and applications to biometric problems.— Biometrics, 1975, vol. 31, p. 545—554.
245. Reinsch C. H. Smoothing by spline fuction.— Numer. Math., 1967, vol. 10, p. 177—183.
246. Richardson M., Kuder G. F. Making a rating scale that measures.— Personnel Journal, 1933, vol. 12, p. 36—40.
247. Saporta G. Liaisons entre plusieurs ensembles de variables et codage de donnés qualitatives.— Paris VI: L' université Purre et Marie Curie, doctoral thesis, 1975.— 110 p.
248. Schatzoff M., Fienberg S., Tsao R. Efficient calculation of all possible regressions.— Technometrics, 1968, vol. 10, p. 768—779.
249. Sclove S. L. Improved estimators for coefficients in linear regression. — J. Amer Statist. Assoc., 1968, vol. 63, p. 596—606.
250. SPSS Statistical Package for the Social sciences Second ed.—Mc. Graw — hill book company, 1975.—675 p.
251. Stein C. Multiple regression. — In: Contributions to Probability and Statistics, Essays in honor of Harold Hotelling, Stanford University Press: Palo Alto, Calif, 1960, p. 424—443.
252. Stein C. Inadmissibility of the usual estimator for the mean of multi-variate normal distribution.— Proc. Berkeley Symp. Math. Statist. and. Prob., 1956, vol. 1, p. 197—206.
253. Taylor L. D. Estimation by minimising the sum of absolute errors.— In: Frontiers in econometrics, N. Y., 1974, p. 169—190.
254. TSA (A program for times series).— In: COMPSTAT-82, Proc. in Computational Statist., Statist. Packages and Facilities Presented at COMPSTAT-82, Wien, 1982, p. 69—72.
255. Van Tassel D. BASIC-Pack Statist. Programs for Small Computers.— London: Prentice-Hall, 1981.— 230 p.
256. Wagner H. M. Linear programming techniques for regression analysis.— J. Amer. Statist. Assoc., 1959, vol. 54, p. 206—212.
257. Wilkinson J. H. The classical error analysis for the solution of linear systems.— J. Inst. Math. Appl., 1974, vol. 10, p. 175—180.
258. Wold S. Spline functions in data analysis.— Technometrics, 1974, vol. 16, № 1, p. 1—11.
259. Youngs E. A., Cramer E. M. Some results relevant to choice of sum and sum-of-product algorithms.— Technometrics, 1975, vol. 17, p. 458—467.

АЛФАВИТНО-ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Сергей Артемьевич Айвазян,
Игорь Семенович Енюков,
Лев Дмитриевич Мешалкин

ПРИКЛАДНАЯ СТАТИСТИКА

ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ

Зав. редакцией *Р. А. Казьмина*

Редактор *Л. Н. Вылегжанина*

Мл. редакторы *А. В. Щурова, В. Л. Долгова*

Техн. редакторы *К. К. Букалова, Г. А. Полякова*

Корректоры *Г. В. Хлопцева, Г. А. Башарина,*
Т. Г. Кочеткова и *М. А. Синяговская*

Худож. редактор *М. К. Гуров*

Переплет художника *Н. А. Пашуро*

ИБ № 1544

Сдано в набор 27.04.84. Подписано в печать 19.12.84 А14099. Формат 84×108¹/₃₂. Бум. офс. № 2. Гарнитура «Литературная». Печать высокая. Усл. п. л. 25,62. Усл. кр.-отт. 25,62. Уч.-изд. л. 27,46. Тираж 13 000 экз. Заказ 244. Цена 1 р. 70 к.

Издательство «Финансы и статистика»,
101000, Москва, ул. Чернышевского, 7
Московская типография № 4 Союзполиграфпрома
при Государственном комитете СССР по делам
издательств, полиграфии и книжной торговли
129041, Москва, Б. Переяславская, 46